Capture-recapture Estimation for Conflict Data and Hierarchical Models for Program Impact Evaluation

Capture-recapture Estimation for Conflict Data and Hierarchical Models for Program Impact Evaluation

A dissertation presented

by

Shira Arkin Mitchell

to

The Department of Biostatistics

in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy
in the subject of
Biostatistics

Harvard University
Cambridge, Massachusetts

April 2014

Dissertation Advisor: Professor Brent Coull Shira Arkin Mitchell

Capture-recapture Estimation for Conflict Data and Hierarchical Models for Program Impact Evaluation

Abstract

A relatively recent increase in the popularity of evidence-based activism has created a higher demand for statisticians to work on human rights and economic development projects. The statistical challenges of revealing patterns of violence in armed conflict require efficient use of the data, and careful consideration of the implications of modeling decisions on estimates. Impact evaluation of a complex economic development project requires a careful consideration of causality and transparency to donors and beneficiaries. In this dissertation, I compare marginal and conditional models for capture recapture, and develop new hierarchical models that accommodate challenges in data from the armed conflict in Colombia, and more generally, in many other capture recapture settings. Additionally, I propose a study design for a non-randomized impact evaluation of the Millennium Villages Project (MVP), to be carried out during my postdoctoral fellowship. The design includes small area estimation of baseline variables, propensity score matching, and hierarchical models for causal inference.

Contents

List of Figures

List of Tables

To Steve Lagakos, who drew me to this department.

Acknowledgments

Thank you to my incredible cohort of brilliant scholars and generous friends. Thank you to my committee: Al, thank you for pushing me to make graphs clearer and cut fluffy words and sentences, and for introducing me to capture-recapture. Brent, thank you for stressing the importance of interpreting parameters, and for seeing the good news in everything. Alan, thank you for helping me see patterns and encouraging me to make sense of them, you are who I want to be when I grow up. Joe, thank you for introducing me to probability and statistics, the value of simplest nontrivial examples, and the importance of telling stories. Thank you to HRDAG for answering my first email to them, opening up an exciting and fruitful collaboration. Thank you to all my teachers and fellow students for having the patience to answer my never-ending stream of questions. Thank you to the department, particularly Jelena, who patiently tolerated my disorganization and offered invaluable moral support. Thank you to my parents, who showed me that math is *fun*.

1. Introduction

A relatively recent increase in the popularity of evidence-based activism has created a higher demand for statisticians to work on human rights and economic development projects. The statistical challenges of revealing patterns of violence in armed conflict require efficient use of the data, and careful consideration of the implications of modeling decisions on estimates. Impact evaluation of a complex economic development project requires a careful consideration of causality and transparency to donors and beneficiaries.

Underreporting the level of violence in armed conflict obscures the true nature of the conflict, precluding development of effective solutions. Violence hidden from official reports and the press endangers the peace process by failing to hold perpetrators and policymakers accountable. Thanks to heroic efforts by journalists and organizations, there are lists of victims for many armed conflicts going on today. With at least two lists recorded from the same population, capture-recapture methods from ecology can be used to estimate the total number of victims. Challenges in obtaining estimates for the level of violence include: sparse data, different models fit the observed data equally well but give substantively different estimates, population heterogeneity, and complex cooperations between organizations collecting data.

For example, since 1964, the Colombian armed conflict has produced tens of thousands of victims. There is great interest in the level and patterns of violence, but the patterns revealed by different sources are categorically different. There is also interest in evaluating the effect of policies such as paramilitary demobilization, which was rolled out department by department, possibly enabling identification of the causal effect, but not if we don't have reliable estimates of the level of violence. The problem is not unique to conflict data, applications in public health and ecology (for Disease Monitoring and Forecasting., 1995; Chao et al., 2001), and estimation of census undercount (Zaslavsky and Wolfgang, 1990, 1993) rely on capture-recapture modeling. Conflict data tends to be particularly messy, because in years and regions with the most violence, there is often the least data, and the organizations collecting data often share information. Additionally, the estimates are highly politicized, so they need to be defensible against scrutiny from

different sides.

Almost equally politicized are evaluations of large-scale economic development projects. The Millennium Villages Project (MVP) is a particularly controversial economic development project, which sprang out of the UN Millennium Summit, the largest gathering of world leaders in history. It is a village-level intervention in 12 clusters of villages across sub-Saharan Africa. Total project cost is $100's of millions, with high-profile donors like George Soros, film stars, and rockstars. The project was not designed with rigorous impact evaluation in mind, so there was no data collected in control sites, and no randomization to treatment. Beyond the usual challenges of causal inference for observational studies (Rubin, 1978; Dehejia and Wahba, 1999; Dehejia, 2005; Imbens and Rubin, 2014), this evaluation presents the additional challenge of estimating baseline data for candidate control sites. We propose to accomplish this through small area estimation (Ghosh and Rao, 1994; Ghosh and Natarajan, 1999; Nadram, 2000; Rao, 2003; Jiang and Lahiri, 2006), combining Demographic and Health Surveys (DHS) data with country censuses. These results will be interesting in their own right, independent of any impact evaluation. Subsequently, fitting the propensity score and causal models will require modification to account for this additional uncertainty.

In this dissertation, I compare existing models for capture recapture, and develop new models that extend the flexibility of existing models to accommodate challenges in the Colombian data, and more generally, in many other capture recapture settings. Additionally, I propose a study design for the impact evaluation of the MVP, to be carried out during my postdoctoral fellowship.

In Chapter 2, we show that collapsed across years, estimates of total killings in the Colombia data differ whether we choose to model *marginal* reporting probabilities and odds ratios, versus modeling the full reporting pattern in a *conditional* (log-linear) model. We use a simulation study to compare marginal and conditional models, generating data from a latent Gaussian model. In Chapter 3, we develop hierarchical log-linear capture-recapture

models that partially pool across time to get yearly estimates for the Colombia data. We investigate two methods to handle groups actively collecting data in different but overlapping time-periods. One imputes the inactive periods, the other uses a mixture model to relax exchangeability assumptions. In Chapter 4, we propose a design for the MVP evaluation, including small area estimation models, selection of control sites, and several candidate causal models. We also perform a power analysis that considers Type S error, the probability that the estimated treatment effect has the incorrect sign, if it is statistically significant, and Type M error, the expected absolute value of the estimate divided by the true effect size, if it is statistically significant (Gelman and Carlin, 2013).

2. A comparison of marginal and conditional models for capture-recapture data with application to human rights violations data

[1]Shira Mitchell, [1,4,5]Al Ozonoff, [2]Alan M. Zaslavsky, [3]Bethany Hedt-Gauthier, [6]Kristian Lum, and [1]Brent A. Coull

[1]Department of Biostatistics, Harvard School of Public Health
[2]Department of Health Care Policy, Harvard Medical School
[3]Department of Global Health & Social Medicine, Harvard Medical School
[4]Clinical Research Center, Boston Childrens Hospital
[5]Department of Pediatrics, Harvard Medical School
[6]Network Dynamics and Simulation Science Laboratory, Virginia Tech

Abstract

Human rights data presents challenges for capture-recapture methodology. Lists of violent acts provided by many different groups create large, sparse tables of data for which saturated models are difficult to fit and for which simple models may be misspecified. We analyze data on killings and disappearances in Casanare, Colombia during years 1998 to 2007. Our estimates differ whether we choose to model *marginal* reporting probabilities and odds ratios, versus modeling the full reporting pattern in a *conditional* (log-linear) model. With 2629 observed killings, a marginal model we consider estimates over 9000 killings, while conditional models we consider estimate 6000-7000 killings. The latter agree with previous estimates, also from a conditional model. We see a two-fold difference between the high sample coverage estimate of over 10,000 killings and low sample coverage lower bound estimate of 5200 killings. We use a simulation study to compare marginal and conditional models with at most two-way interactions and sample coverage estimators. The simulation results together with model selection criteria lead us to believe the previous estimates of total killings in Casanare may have been biased downward, suggesting that the violence was worse than previously thought. Model specification is an important consideration when interpreting population estimates from capture recapture analysis and the Casanare data is a protypical example of how that manifests.

2.1 Introduction

Since 1964, the Colombian armed conflict between the military, guerrilla, and paramilitary groups has killed tens of thousands of people, and displaced millions. Underreporting the level of violence obscures the true nature of the conflict, precluding development of effective solutions. Violence hidden from official reports and the press endangers the peace process by failing to hold perpetrators and policy-makers accountable.

Both government and nongovernment groups (NGOs) in Colombia report killings and

disappearances. If we assume that a documented case from any list truly happened, then no single list from the government or NGO is complete. In this paper, we use the statistical technique of *capture-recapture* to estimate the number of killings and disappearances in the Casanare region of Colombia over the years 1998 to 2007, using data provided by 15 groups. Casanare is in the central eastern region of Colombia with a population of 300,000. It contains oil fields and a British Petroleum pipeline. Injection of cash into the economy from oil profits without government capacity for managing order created an environment conducive to violence by guerrilla and paramilitary groups (Davy et al., 1999). Human rights groups and policy-makers ask: How many killings and disappearances occurred in Casanare?

With its basis in ecology at the turn of the twentieth century, capture-recapture (also known as multiple systems estimation) has also been used to estimate totals for human populations. Early work using capture-recapture for human rights data was done by HRDAG, the Human Rights Data Analysis Group, part of the Benetech Initiative. HRDAG used capture-recapture to estimate the number of killings and disappearances in Casanare (Lum et al., 2010; Guberek et al., 2010), but the challenges of these data motivates further investigation.

We fit various capture-recapture models to killings and disappearances data from Casanare. Our estimates of total disappearances remain mostly stable, but estimates of killings vary with model choice. We see large difference between estimates from modeling *marginal* reporting probabilities of each list and *marginal* odds ratios between lists, versus modeling the full reporting pattern, with parameters describing reporting probability of a list *conditional* on other lists. These two model structures are referred to in the literature as *marginal* and *conditional (log-linear)* models. With 2629 observed killings, a marginal model estimates over 9000 killings, while conditional models estimate 6000-7000 killings. The latter agree with the HRDAG estimates, also obtained from a conditional model. We also compare the estimates from the marginal and conditional models to the sample coverage approach (Chao and Tsay, 1998; Tsay and Chao, 2001; Chao et al.,

2001). We see a two-fold difference between the high sample coverage estimate (HSC) of over 10,000 killings and low sample coverage lower bound estimate (LSC) of 5200 killings. The standard error from the HSC exceeds one-third of the population size estimate, suggesting there may not be enough information to accurately estimate the total number of killings (Chao et al., 2001). The question is whether one should put more faith into the estimates from previous reports in Lum et al. (2010), or larger estimates that suggest these initial figures are underestimates of the number of killings.

The literature provides few systematic comparisons between marginal and conditional models. Glonek and McCullagh (1995) prove parameter orthogonality results in the marginal model that suggest that misspecification of higher-order parameters would not bias estimates of lower-order parameters. However, in the case of capture-recapture, where interest focuses on prediction of missing counts, orthogonality of marginal parameters does not necessarily translate to population size estimates that are robust to model misspecification.

Chao et al. (2001) compare ecological models, which model the probability of capturing animal i in list j, to log-linear models that model the full capture pattern, (Chao et al., 2001). Ecological models are marginal with respect to the lists, but they do not incorporate list dependence (other than dependence induced by heterogeneity of capture probability). The Rasch model is a type of ecological model, (Darroch et al., 1993; Coull and Agresti, 1999; Agresti, 1994). Chao et al. (2001) recommend ecological models for animal studies, where captures are independent and there are over four captures. They recommend log-linear models for epidemiology, when two to four possibly dependent lists are available. There is a connection between the two approaches. The Rasch model is equivalent to the quasi-symmetric log-linear model with some moment constraints (Darroch et al., 1993). The generalized Rasch model is equivalent to the partial quasi-symmetric log-linear model. In our paper, we have fit both quasi-symmetric and partial quasi-symmetric log-linear models.

Bartolucci and Forcina (2001, 2006) use marginal models that extend the Rasch model to include two-way list associations. They include latent classes for unobserved heterogeneity of capture, observed covariates, and marginal interactions between lists. Choosing a conditional model instead, E Stanghellini (2004) consider latent classes, observed covariates, and log-linear interactions between covariates, lists, and latent class. Both fit their models to data counting diabetic patients in a town in northern Italy (four lists) and obtain similar estimates.

For a saturated model, marginal and conditional formulations are re-parametrizations of each other. However, in situations with at least four lists, we often do not fit a model with many higher order interactions. Zero counts cause fitting difficulty, and confidence intervals are wide. It is common to restrict focus to models with two-way interactions between lists, eliminating higher-order interactions, or to fit quasi-symmetric models where heterogeneous two-way interactions are described by a single parameter (Chao et al., 2001; for Disease Monitoring and Forecasting., 1995; Agresti, 1994). These simplifications make the marginal and conditional models different.

In Section 3.2 we describe the motivating dataset of violence records in Casanare. In Section 3.3 we describe candidate models we will compare. Section 2.4 presents results from marginal and conditional model analyses of the motivating Casanare data. In Section 2.5 we describe simulations to compare models. We then make conclusions.

2.2 Motivating Dataset - Casanare

Our data are lists of violent events, both killings and disappearances, provided by 15 groups, both government and NGOs. These lists are called *sources* or *captures* in the capture-recapture literature. After de-duplication of records, there are 2629 reported killings and 872 disappearances. The seven longest lists for killings combined contain all 2629 observed killings, and the seven longest lists for disappearances contain 867 dis-

appearances, missing only five. Some groups specialize in reporting killings or disappearances, with three groups providing lists in both the top seven for killings and disappearances. For killings, three of the seven lists are from NGOs and for disappearances two are from NGOs. A majority of records appear in only one list: 500 of the disappearance records and 1847 of the killings records. For information about the groups, the matching algorithm to connect observations across lists, and raw data descriptives, see Appendix A.2.2.

Many zero cells in a table cross-classifying lists causes large standard errors and unstable results when fitting models (see Agresti, 2002, p.394). To alleviate this sparsity and because the longest lists contain most of the observed records, we take only the top seven lists for each violation type (killings and disappearances).

2.3 Candidate models

We wish to estimate the size, N, of a closed population of killings or disappearances, using lists of violent events provided by J sources. Let n be the number of events recorded in at least one list. Let $n_{\mathbf{k}}$ be the number of events with recording pattern $\mathbf{k}$, a string of 1's denoting recording in a list and 0's denoting non-recording, of length J. These are cell counts in a 2^J contingency table cross-classifying the lists. We assume a multinomial sampling plan $\mathbf{n} = \{n_{\mathbf{k}}\} \sim Multinomial(N, \{\pi_{\mathbf{k}}\})$ with $\boldsymbol{\pi} = \{\pi_{\mathbf{k}}\}$ the cell probabilities. The observed data are $2^J - 1$ cell counts with sum n, where we do not observe the cell with events not recorded by any list. The problem is to predict this missing cell count, $n_{\mathbf{o}}$, or equivalently, estimate N, the sum of all counts in the table. The traditional approach to this problem fits a model to the $2^J - 1$ observed cells, and predicts the unobserved cell using the fitted value for that cell from the model fit.

The class of *generalized log-linear models* (GLLMs) represents a large class of models, including both conditional and marginal models, and has the form $L(\boldsymbol{\pi}) = \mathbf{C} \log \mathbf{A}\boldsymbol{\pi} = \mathbf{X}\boldsymbol{\beta}$

(McCullagh and Nelder, 1989; Glonek and McCullagh, 1995; Glonek, 1996). A conditional model with heterogeneous two-way conditional interactions has matrices $\mathbf{C}$ and $\mathbf{A}$ as identity, and

$$\log \pi_{\mathbf{k}} = \lambda_0 + \sum_j \lambda_j k_j + \sum_{j \neq j'} \lambda_{j,j'} k_j k_{j'} \tag{1C}$$

where $\lambda_{j,j'}$ is the log-odds ratio of recording in lists j and j' conditional on recording pattern in other lists. For a marginal model, matrix $\mathbf{A}$ selects margins from $\boldsymbol{\pi}$ and matrix $\mathbf{C}$ sets up contrasts to create marginal log-odds and log-odds ratios. A marginal model with heterogeneous two-way marginal interactions is

$$\log(\eta^j) = \beta_j, \quad \log(\eta^{j,j'}) \ = \omega_{j,j'}, \tag{1M}$$

where η^j is the marginal odds of recording in list j, and $\eta^{j,j'}$ is the marginal odds ratio of recording in lists j and j'. In both models 1C and 1M we also have parameter N, so each has $1 + J + \binom{J}{2}$ parameters.

We note that in model 1M, parameters β_j, $\omega_{j,j'}$ are not *variation independent*, defined such that the range of values for one does not depend on the other's value (Bergsma and Rudas, 2002). Thus, there may exist values for them that do not give a joint distribution on $n_{\mathbf{k}}$. However, reasonable values for parameters yield full distributions, seen by a positive-definite variance matrix for parameter estimates, and fitted probabilities within $[0, 1]$, (Molenberghs and Lesaffre, 1999). The log-linear parametrization is variation independent.

Candidate models include marginal and conditional model pairs of three types:

- Heterogeneous two-way marginal and conditional interaction models 1C and 1M.
- Homogeneous two-way conditional and marginal interaction models, also referred to as two-way *quasi-symmetry (QS)* models

$$\log \pi_{\mathbf{k}} = \lambda_0 + \sum_j \lambda_j k_j + \sum_{j \neq j'} \lambda k_j k_{j'}, \tag{2C}$$

$$\log(\eta^j) = \beta_j, \quad \log(\eta^{j,j'}) \ = \omega. \tag{2M}$$

In both models 2C and the corresponding 2M we have $1 + J + 1$ parameters.

- Zero cell counts in the Casanare data cause parameter estimates tending to infinity when fitting models 1M and 1C. The *three-level quasi-symmetry models (QS3)* are motivated by HRDAG's suggestion that dependence is lowest between lists of different types. We restrict interactions between two NGOs to be equal, between two government lists to be equal, and between a government and NGO to be equal:

$$\begin{aligned}\log \pi_{\mathbf{k}} = \lambda_0 + \sum_j \lambda_j k_j + \\ \sum_{j,j' \in NGOs} \lambda_{NGO} k_j k_{j'} + \sum_{j,j' \in govt} \lambda_{govt} k_j k_{j'} + \sum_{j \in NGOs, j' \in govt} \lambda_{mix} k_j k_{j'},\end{aligned} \tag{3C}$$

$$\log(\eta^j) = \beta_j, \quad \log(\eta^{j,j'}) \ = \begin{cases} \omega_{NGO} & j, j' \in NGOs \\ \omega_{govt} & j, j' \in govt \\ \omega_{mix} & j \in NGOs, j' \in govt. \end{cases} \tag{3M}$$

To fit both marginal and conditional models, we use Joseph Lang's `R` program `mph.fit`, (Lang, 2004, 2005). See Appendix A.1.1 for details on model fitting algorithms.

The *sample coverage approach*, proposed by Chao and Tsay (1998) and Tsay and Chao (2001) uses a measure to quantify list overlap information in order to estimate the missing cell count. We use two sample coverage estimators. As in Chao et al. (2001), we define

$$S_j \equiv \sum_{\substack{k_j=1, \\ k_m=0 \ \forall m \neq j}} n_{\mathbf{k}} = \text{number of events in list } j \text{ only},$$

$$T_j \equiv \sum_{k_j=1} n_{\mathbf{k}} = \text{number of events in list } j,$$

$$A(i,j) \equiv \sum_{\substack{k_i=1, \\ k_m=0 \ \forall m \neq i,j}} n_{\mathbf{k}} + \sum_{\substack{k_j=1, \\ k_m=0 \ \forall m \neq i,j}} n_{\mathbf{k}},$$

$$B(i,j) \equiv \sum_{\substack{k_i=k_j=1, \\ k_m=0 \ \forall m \neq i,j}} n_{\mathbf{k}}, \text{ and } D \equiv n - \frac{1}{J} \sum_{j=1}^{J} S_j.$$

An estimator for the sample coverage is,

$$\widehat{C} = 1 - \frac{1}{J}\sum_{j=1}^{J}\frac{S_j}{T_j}.$$

Then when list overlap is large enough, the high sample coverage (HSC) estimator is

$$\widehat{N} = \left[\frac{D}{\widehat{C}} - \frac{1}{t\widehat{C}}\sum_{i<j}A(i,j)\right]\left\{1 - \frac{1}{t\widehat{C}}\sum_{i<j}\frac{A(i,j)B(i,j)}{T_iT_j}\right\}^{-1}.$$

For relatively low sample coverage data, the low sample coverage (LSC) estimator is

$$\widehat{N} = \frac{D}{\widehat{C}} + \frac{1}{t\widehat{C}}\sum_{i<j}A(i,j)\widehat{\gamma}_{ij},$$

where

$$\widehat{\gamma}_{ij} \equiv \frac{B(i,j)}{T_iT_j}\left[\frac{D}{\widehat{C}} + \frac{1}{t\widehat{C}}\sum_{r<s}A(r,s)\left(\frac{D}{\widehat{C}}\frac{B(r,s)}{T_rT_s} - 1\right)\right] - 1.$$

2.4 Casanare Data Analysis

Population estimates for the Casanare data are presented in Table 2.1, including estimates from fitting models 2M, 2C, 3M, and 3C and the LSC and HSC estimators. Interaction parameter estimates and goodness of fit statistics are available in Appendix A.1.2. Reflecting well-known collaboration among NGO groups, both marginal and conditional log odds ratios for recording killings are high among NGO lists. Log odds ratios reflecting interactions between government lists and NGO lists are lower for disappearances. Log odds ratios reflecting interactions between a government and NGO list are much lower than log odds ratios reflecting interactions between two lists of the same type. Due to high association among NGOs, we explore the effect of collapsing NGO lists, treating all events recorded by NGOs as coming from a single source: $k_{NGOs} \equiv \vee_{j\in NGOs}k_j$. For killings (disappearances) we collapse the three (two) NGOs, so we fit the QS models 2C and 2M with $J = 5$ ($J = 4$) lists:

$$\log\pi_{\mathbf{k}} = \lambda_0 + \sum_{j\in govt}\lambda_j k_j + \lambda_J k_{NGOs} + \sum_{j\neq j'}\lambda k_j k_{j'}, \tag{4C}$$

$$\log(\eta^j) = \beta_j, \quad \log(\eta^{j,j'}) \quad = \omega, \tag{4M}$$

as well as the *two-level quasi-symmetry models (QS2)*

$$\log \pi_{\mathbf{k}} = \lambda_0 + \sum_{j \in govt} \lambda_j k_j + \lambda_J k_{NGOs} + \sum_{j,j' \in govt} \lambda_{govt} k_j k_{j'} + \sum_{j' \in govt} \lambda_{mix} k_{j'} k_{NGOs}, \tag{5C}$$

$$\log(\eta^j) = \beta_j, \quad \log(\eta^{j,j'}) \quad = \begin{cases} \omega_{govt} & j, j' \in govt \\ \omega_{mix} & j = NGOs, j' \in govt. \end{cases} \tag{5M}$$

Joseph Lang's `R` program `mph.fit` reports failure to converge in fitting the heterogenous two-way marginal and conditional models 1M and 1C to the Casanare data. However, alternative software exists to fit log-linear models, such as `glm` in `R`. The `glm` function does report convergence, but the flat deviance profile shown in Figure 2.1 shows that the resulting estimate is unstable due to a flat likelihood for $N > 10,000$. We use `mph.fit` to compare corresponding marginal and conditional models for consistency.

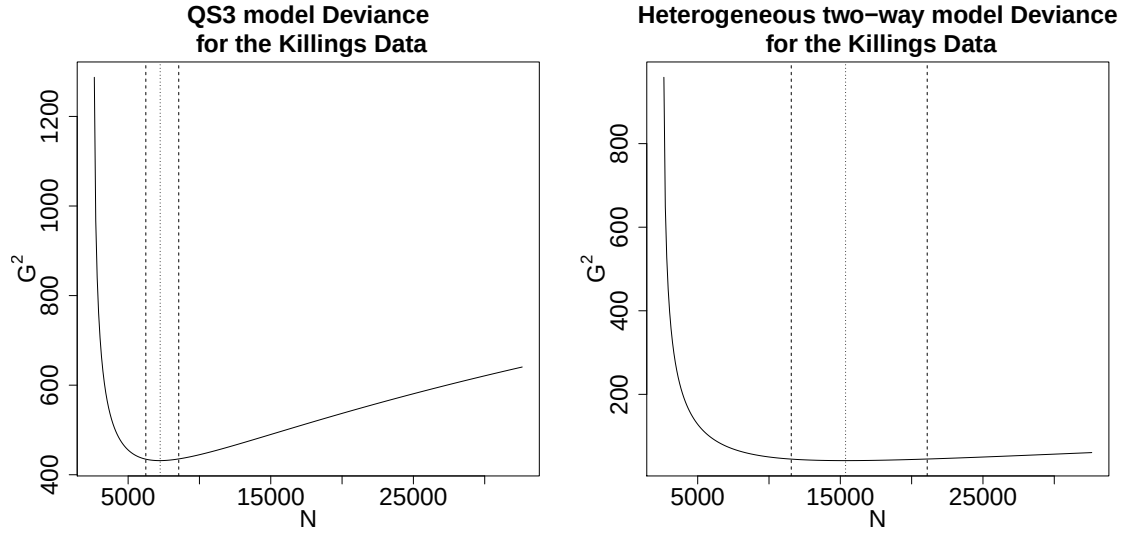


Figure 2.1: Plots show the deviance (G^2) profile for N for the killings data, fitting the QS3 conditional model 3C and the heterogeneous two-ways conditional model 1C. Vertical dotted lines show the maximum likelihood estimate and confidence limits for N.

We estimate list completeness values by taking the number of records in each list and dividing by the maximum estimate of N across models. List completeness values were

estimated to be 1% to 21% for killings and 4% to 18% for disappearances, where both had lower list completeness for NGOs.

For killings, the marginal models give higher estimates of totals than the corresponding conditional models. This difference is smallest for the marginal model 2M $\widehat{N} = 6142(5447, 7265)$ versus conditional model 2C $\widehat{N} = 5689(5174, 6356)$, models with homogeneous two-way interactions. The difference is largest for the marginal model 3M $\widehat{N} = 9277(7538, 11720)$ versus conditional model 3C $\widehat{N} = 6751(5902, 7811)$, models that allow the two-way interactions to depend on list type. The heterogeneous two-way model 1C estimates $\widehat{N} = 14334(10851, 19292)$ killings. HRDAG estimates $\widehat{N} = 6215(3944, 9983)$ killings.

For disappearances, estimates from marginal and conditional models are similar with confidence intervals largely overlapping. The marginal model 3M gives $\widehat{N} = 1382(1249, 1552)$ versus conditional model 3C $\widehat{N} = 1508(1350, 1706)$, comparing models that allow the two-way interactions to depend on list type. The heterogeneous two-way model 1C estimates $\widehat{N} = 1940(1413, 2976)$ disappearances. HRDAG estimates $\widehat{N} = 2653(1270, 5552)$ disappearances.

In all but one case (disappearances, models that collapse NGOs, 4M versus 5M) a likelihood ratio test rejects the homogeneous two-way interactions model in favor of a model that allows the two-way interactions to depend on list type. For the conditional models, likelihood ratio tests reject the QS3 models that allow two-way interactions to depend on list type in favor of the heterogeneous two-way interaction models. The marginal and conditional model pairs have the same number of parameters, so we can compare the deviance (G^2) from each model. For the killings data, the marginal models have lower deviance than the corresponding conditional models. For the disappearances data, the conditional models have lower deviance (see Appendix A.1.2). Likelihood ratio tests and comparing deviances reflects how well the models fit to the observed data, not necessarily how well they predict the missing cell.

The sample coverage approach gives slightly lower estimates for disappearances than our marginal and conditional models, with the high sample coverage and low sample coverage approaches giving similar estimates. For killings, the high sample coverage approach estimates over 10,000 killings, but with a bootstrap standard error of 4298. The sample coverage is estimated to be 58%. Chao et al. (2001) recommend the sample coverage to be at least 55% and that the bootstrap standard error not exceed one-third the estimated total in order to use the high sample coverage estimator. Though the sample coverage meets the threshold, the high standard error suggests there may not be enough information in the data to estimate the population size. Thus, we may prefer the lower-bound estimator for low sample coverage (LSC), which estimates roughly 5200.

2.4.1 Relation of Findings to Existing Results

HRDAG takes three-list subsets of the 15 lists and considers *graphical models* on the three lists. Graphical models are a subset of log-linear models where conditional dependence relationships can be represented graphically (Madigan and York, 1997; Darroch et al., 1980; Dawid and Lauritzen, 1993). HRDAG uses Bayesian model averaging to combine estimates (Lum et al., 2010). The comparability of these three-list subsets is not obvious because discarding records not captured by a subset results in different data for each subset. HRDAG estimates 2653 disappearances with 95% credible interval (1270, 5552), and 6215 killings with 95% credible interval (3944, 9983).

HRDAG's estimate for total killings is most similar to our conditional model results, or the results from the marginal model when we fit the QS model 2M. Their point estimate of 6215 lies outside confidence intervals from models 3M, 4M, 5M. Their credible interval lower bound for total number of disappearances is roughly equal to our confidence interval lower bounds for the disappearance data, but their point estimate and upper bound are higher than all of our point estimates and upper bounds.

2.5 Simulation Study

In order to explore when marginal and conditional model population estimates differ, and how they perform in terms of bias and variance, we conducted a simulation study. We simulate data by bisecting a multivariate normal distribution to produce dichotomous observed values indicating recording or non-recording on a list. This *latent Gaussian model* is the *tetrachoric correlation model* if we generate a bivariate statistic. We broaden the term to our setting, with more than two lists, referring to our generating model as the "tetrachoric model." We choose this model so that neither the conditional nor marginal models are the correct model, and we have a formulation that allows us to flexibly examine different association structures among lists. The tetrachoric model produces data with non-zero three-way odds ratios, as we will see below. For J lists, we generate N vectors $(Y_1, \ldots, Y_J)' \sim N_J(\boldsymbol{\theta}, \boldsymbol{\Sigma})$, where $\boldsymbol{\theta} = (\theta_1, ..., \theta_J)'$, $k_j = 1$ if $Y_j \geq 0$, for $j = 1, \ldots, J$, which gives data $\{n_{\mathbf{k}}\}$, the number of events with recording pattern $\mathbf{k} = \{k_j\}$. We compute $E[n_{\mathbf{k}}] = \mu_{\mathbf{k}}$ using the multivariate normal distribution. Let $\mathbf{p} = (p_1, ..., p_J)$ where p_j is the probability of recording an event on list j, i.e. the completeness of list j. We perform simulations of two broad types. We do base simulations, taking the simplest correlation structure: $\boldsymbol{\Sigma}$ exchangeable, where a parameter d describes inter-list correlation, the same between any pair of lists. Next, motivated by the Casanare data, we look at a block structure, with blocks for NGOs and government groups. For example, with 3 NGOs and 4 government groups, we have

$$\boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{A} & \mathbf{C} \\ \mathbf{C}' & \mathbf{B} \end{bmatrix},$$

where $\mathbf{A} = \mathbf{1}_3\mathbf{1}_3'a - (1-a)\mathbf{I}_3$, $\mathbf{B} = \mathbf{1}_4\mathbf{1}_4'b - (1-b)\mathbf{I}_4$, $\mathbf{C} = \mathbf{1}_3\mathbf{1}_4'c$, b is government list association, a is NGO association, and c is association across type. We take b as 0.2 or 0.65, range a from 0.5 to 0.95, and set $c = b/2$. We vary the number of NGOs to be 2 or 3 (5 or 4 government lists), to best approximate the Casanare data.

We fix population size $N = 2000$. For the base simulation scenario, we take $J = 4$ and $J = 6$ to examine trend across number of lists. For Casanare-inspired simulations we take

$J = 7$. We examine four different ranges of values for list completeness.

The first three are characterized by average list completeness: low (0.08-0.38), medium (0.27-0.73), or high (0.62-0.92). These ranges roughly correspond to different settings in which capture-recapture is applied. In census settings, completeness may be high. For human rights data, violent events are very difficult to document. In ecology, recording probabilities vary according to sampling fraction. We also look at the scenario where one "master" list attempts census with a high (0.88) completeness, and the remaining $J - 1$ lists have low completeness (0.08-0.38). We term this "varied" completeness. For Casanare-inspired simulations, we take very low values for list completeness (0.03-0.18) with the lower list completeness for NGOs. We generate 40 datasets for each combination of number of lists, completeness level, and $\mathbf{\Sigma}$. For details on simulation conditions, see Appendix A.1.3.

For the base simulations we fit models 1M, 1C, 2M, and 2C. For the Casanare simulations we fit models 2M, 2C, 3M, and 3C. For both we also use the sample coverage approach, and compute both the high sample coverage (HSC) and low sample coverage (LSC) estimates.

2.5.1 Results - base simulations

To characterize the simulated data from the tetrachoric model, we plot information about cell means $\{\mu_{\mathbf{k}}\}$ in supplementary Figures A.1 and A.2. All models we fit assume no three-way and higher marginal or conditional interactions. The QS models 2M and 2C assume homogeneous two-way ORs. For d values further from zero, magnitudes of three-way ORs are higher and there is more variation in two-way ORs. The mean cell count for the missing cell ($\mu_{\mathbf{o}}$) is larger when the lists are less complete or association between lists increases.

In Figure 2.2 we plot $\widehat{N} - N$ for the estimates from marginal model 2M and conditional

model 2C in row one, and $\widehat{N} - N$ for the low and high sample coverage estimates in row two. In row three we plot the root mean square error (RMSE) for the marginal and conditional model estimates, and in row four the RMSEs for the low and high sample coverage estimates. In supplementary Figures A.7, A.8, and A.9 we provide results from fitting the two-way models 1M and 1C as well as results from simulations with $J = 4$ lists. Patterns were similar for $J = 4$ and $J = 6$ lists. The marginal models are less biased than the conditional models for positive d values and low completeness, with the conditional models tending to underestimate N more. For d near zero and negative, and for medium completeness, the models perform similarly, except at high d, where again the conditional models tend to underestimate N more than the marginal models. For high completeness the models perform similarly except for high d, where the two-way marginal model 1M overestimates N, and the QS conditional model 2C overestimates N. For varied completeness, estimates from the two models are similar, with the marginal models having slightly less bias at high d.

For low and medium completenesses, marginal models 1M and 2M are less negatively biased than conditional models 1C and 2C when d is higher. For higher d, marginal and conditional models' estimates are more different, and both are negatively biased. We see that when estimates from marginal and conditional models are different, both are negatively biased. The variances of the estimates from the marginal and conditional models are similar. Almost all of the RMSE comes from bias rather than variance.

Supplementary Figure A.10 plots coverage of 95% profile likelihood intervals for the base simulations. A 95% profile confidence interval is $n + \{n_{\mathbf{o}}\}$ such that $G^2(n_{\mathbf{o}}) - G^2(\widehat{n}_{\mathbf{o}}) \leq \chi^2_{1,0.05}$, where $G^2(n_{\mathbf{o}})$ is the deviance statistic for the model if the missing cell were $n_{\mathbf{o}}$, and $\widehat{n}_{\mathbf{o}}$ is the maximum likelihood estimate, which minimizes the deviance (Cormack, 1992). Coverage is low for high values of d, especially for low and high completeness. The marginal model has coverage much closer to 95% for the QS model 2M relative to 2C, especially for low and high completeness. The conditional model has slightly better coverage for the heterogeneous two-way model 1C relative to 1M for high completeness

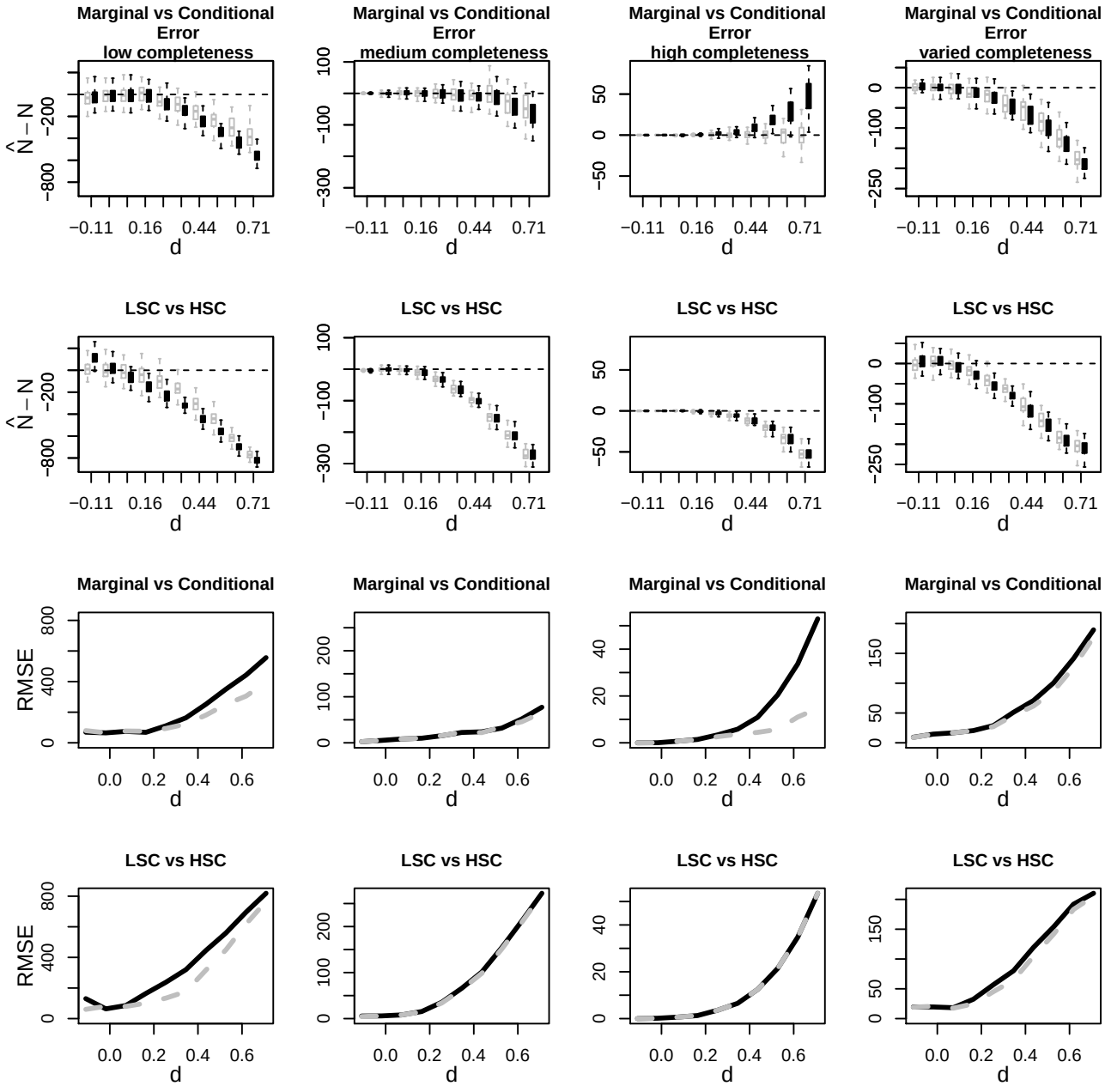


Figure 2.2: Results from the base simulations: $N = 2000$ true total population size, $J = 6$ lists, exchangeable correlation structure. In rows one and two, we plot the distribution of $\widehat{N} - N$ across simulations (as boxplots), and in rows three and four we plot the RMSE, where we note that almost all the MSE is attributable to the bias rather than variance. We use solid lines and filled black boxplots for QS conditional model 2C and the low sample coverage estimator, and dashed lines and empty gray boxplots for the QS marginal model 2M and high sample coverage estimator.

only.

The low and high sample coverage estimates are similar to each other, except for the low completeness scenario. We note that when d is negative, the low sample coverage estimate serves as an upper bound, but when d is positive, it serves as a lower bound. The high sample coverage estimates tend to be more biased downward than the marginal or conditional estimates, and the low sample coverage estimates give a very low lower bound for high d.

2.5.2 Results - Casanare simulations

Figure 2.3 shows $|\widehat{N}_M - N| - |\widehat{N}_C - N|$ and $\widehat{N}_M - \widehat{N}_C$ for estimates of N from the marginal QS3 model 3M and the conditional QS3 model 3C, for different numbers of NGOs, and different levels of association between NGOs and government groups. Parameter b is government list association, a is NGO association, and c is association across type, which we set at $c = b/2$. We provide results from fitting QS models 2M and 2C in supplementary Figure A.12. We plot ω_{NGO}, ω_{govt}, and ω_{mix} from model 3M fit to the $\{\mu_{\mathbf{k}}\}$ of the simulated tetrachoric data (solid lines). Dotted (dashed) lines show these parameters fit to the real killings (disappearances) data. These lines are horizontal, because the axis changes the a parameter for simulating data, and these lines mark fits to the real data that do not change with simulation conditions. Values plotted appear in Table A.1 in the Appendix. For finding which simulation conditions are most pertinent to the real data, we look where solid lines cross dotted (for killings) or dashed (for disappearances). The rightmost two boxplots within the rightmost column, when both a and b are high and there are three NGOs, correspond most closely to killings. The leftmost boxplot, with lower a and b, corresponds most closely to disappearances. In the bottom row of Figure 2.3 we plot the mean cell count for the missing cell ($\mu_{\mathbf{o}}$), which is larger when association between lists increases, either increasing a or increasing b.

In Figure 2.4 we plot the errors $\widehat{N} - N$ for the marginal and conditional model estimates,

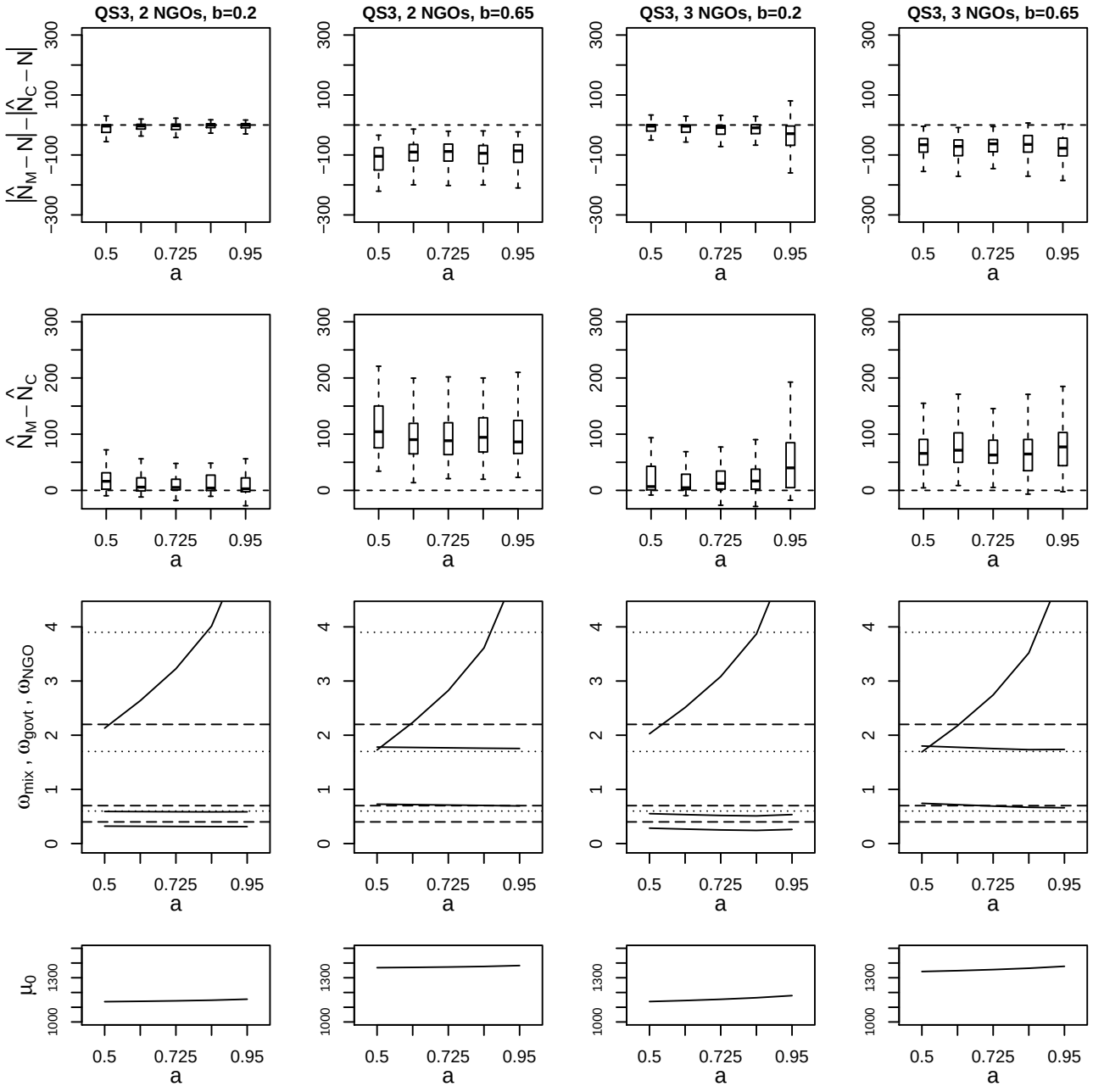


Figure 2.3: The first two rows of the plot show $|\widehat{N}_M - N| - |\widehat{N}_C - N|$ and $\widehat{N}_M - \widehat{N}_C$ for estimates of N from the marginal QS3 model 3M versus the conditional QS3 model 3C fit to simulated data. Data simulated have $N = 2000$, $J = 7$ lists, and a block tetrachoric correlation structure by list type, where b is government list association, a is NGO association, and c is association across type, which we set at $c = b/2$. The third row shows ω_{NGO}, ω_{govt}, and ω_{mix} from model 3M fit to the $\{\mu_{\mathbf{k}}\}$ of the simulated data (solid lines). Overlaid are dotted (dashed) lines showing these parameters fit to the killings (disappearances) data. We plot the mean cell count for the missing cell ($\mu_{\mathbf{o}}$) in the bottom row.

for the low and high sample coverage estimates. The layout of the columns is identical to Figure 2.3, so we can examine the simulation conditions corresponding to the killings and disappearances data in the rightmost and leftmost boxplots respectively.

Marginal models 2M and 3M generally give higher estimates of N than the conditional models 2C and 3C. This difference increases with higher association between government groups, parameter b. Comparing columns one and three, we see that when association between government groups is low ($b = 0.2$), adding an NGO in place of a government group increases the difference between marginal and conditional models. In the base scenario, we saw that for low list completeness and high tetrachoric correlation, the QS marginal model 2M reports a higher estimate of N than the QS conditional model 2C. The pattern from the base simulation holds: higher correlation increases difference between estimates from marginal and conditional models. Increasing association between NGOs via parameter a does not change the difference much, perhaps because there are only 2 or 3 NGOs.

For the scenario corresponding to killings data, marginal model 3M gives a higher estimate than conditional model 3C by a median approximately 5% of the total $N = 2000$. Looking at Figure 2.4, we see that the conditional QS3 model 3C underestimates by 40%, and the marginal QS3 model 3M by 35%. Thus, we see that for the scenario corresponding to killings data, both estimates are biased downward, with the conditional model's estimate more so. In contrast, for the scenario corresponding to the data on disappearances, the differences between the estimates from marginal and conditional models are less. Both estimates are only slightly biased downward by a median of approximately 6%.

Supplementary Figure A.11 plots coverage of 95% profile likelihood intervals for the Casanare-inspired simulations. For high correlation among NGOs ($b = 0.65$), coverage is near zero, consistent with high correlation in base simulations, and driven by the large negative bias in Figure 2.4. For the scenario corresponding to disappearances, coverage

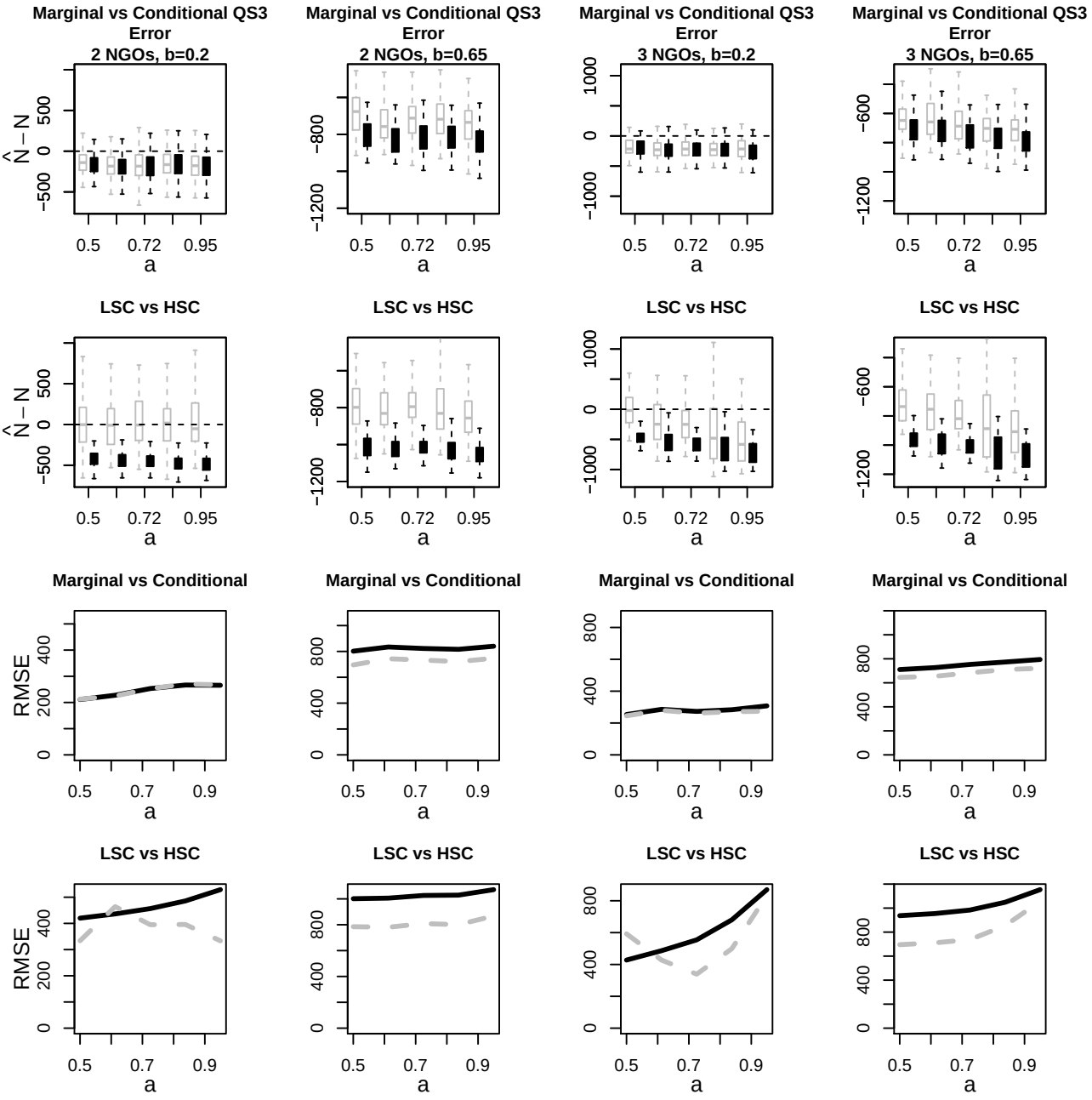


Figure 2.4: In rows one and two, we plot the distribution of $\widehat{N} - N$ across simulations (as boxplots), and in rows three and four we plot the RMSE, where we note that almost all the MSE is attributable to the bias rather than variance. We use solid lines and filled black boxplots for QS3 conditional model 3C and the low sample coverage estimator, and dashed lines and empty gray boxplots for the QS3 marginal model 3M and high sample coverage estimator. Data simulated have $N = 2000$, $J = 7$ lists, and a block tetrachoric correlation structure by list type, where b is government list association, a is NGO association, and c is association across type, which we set at $c = b/2$.

for both models are near 95%.

We see in Figure 2.4 row two that both the sample coverage estimates are underestimates for the scenario corresponding to killings. Because the low sample coverage estimates should be regarded as a lower bound, we conclude these estimates are not tight bounds in these simulated scenarios. For the scenario corresponding to disappearances, the high sample coverage estimator does well in terms of bias.

2.6 Conclusions

Human rights data presents challenges for capture-recapture methodology. The table cross-classifying lists has high dimension when many lists are available. The spread of observed events over many cells results in many zero or near-zero cells, so saturated models give very wide confidence intervals. Models with fewer parameters may be misspecified.

We fit various capture-recapture models to killings and disappearances data from Casanare, Colombia between 1998 and 2007. Our estimates of total disappearances are mostly stable, but with 2629 observed killings, a marginal model estimates over 9000 killings, while conditional models estimate 6000-7000 killings, the latter agreeing with previous estimates (Lum et al., 2010). We see a two-fold difference between the high sample coverage estimate (HSC) of over 10,000 killings and low sample coverage lower bound estimate (LSC) of 5200 killings (Chao and Tsay, 1998; Tsay and Chao, 2001; Chao et al., 2001). The standard error from the HSC exceeds one-third the population size estimate, so there may not be enough information to accurately estimate the total number of killings (Chao et al., 2001). The conditional model with heterogeneous two-way interactions estimates over 14,000 killings, though we caution that the flat deviances at $N > 10,000$ make point estimates unstable, so we focus on the lower bounds of the confidence intervals for this model, which is around 10,000 killings.

Using simulated data from the tetrachoric correlation model, we see that when marginal and conditional model estimates differ, under low list completeness and positively correlated lists, both are biased downward, with the conditional model more so. Our simulations do not show a situation where the marginal model overestimates the total more than the conditional model. Under the tetrachoric model that we simulate with one and two-way margins similar to the Casanare killings data, the marginal model with list type taken into account, model 3M, is less biased downward than the conditional model 3C. If the tetrachoric model is reasonably close to the true distributions of the Casanare data, the simulations suggest that the previous estimates of total number of killings may be too low. The high sample coverage estimate also lead us to believe previous estimates may have been too low. Further investigation is needed to establish analytically why the marginal model estimates are higher for these data.

In data on killings in Casanare between 1998 and 2007, there is high collaboration between NGOs and between government groups. We saw in our simulations that higher associations between lists on the tetrachoric scale result in both marginal and conditional models being biased downward. For disappearances, list collaboration is less, with less difference between estimates from the marginal and conditional models we fit. Our models estimate roughly 1400-1500 disappearances in Casanare between 1998 and 2007.

Based on our analysis of the Casanare data, we recommend incorporating information about the lists (in our example: government or NGO) as a parsimonious way to model list interactions. Due to sparse data, fitting algorithms often fail to converge for a heterogeneous two-way interactions model. We present a useful compromise between quasi-symmetry and a heterogeneous two-way model by incorporating information about the lists.

None of the models considered in this paper, when fit to the killings data, gives qualitatively lower estimates than the HRDAG estimates. Our investigation suggests that the violence was worse than previously thought. This issue of possible underestimates

reinforces our message that model specification is an important consideration when interpreting population estimates from capture recapture analysis.

The Casanare data also include some information about year and location of disappearance or killing. Future research will investigate how estimates of violent acts vary over time and across subregions of Casanare.

Table 2.1: Casanare data results: population total estimates (confidence interval) *[model fit] or estimator used*. For marginal and conditional models, profile confidence intervals and used. For the sample coverage approach, bootstrap confidence intervals are used.

Killings, $n = 2629$							
	QS models		QS2/QS3 models		two-way model	Sample coverage approach	
	Marginal	Conditional	Marginal	Conditional	Conditional	HSC	LSC
3 NGOs, 4 govt	6142 (5447,7265) *[2M]*	5689 (5174,6356) *[2C]*	9277 (7538,11720) *[3M]*	6751 (5902,7811) *[3C]*	14334 (10851, 19292) *[1C]*	10086 (5239, 23934)	5206 (4684, 5862)
Collapsed NGOs, 4 govt	9341 (7447,13296) *[4M]*	5858 (5174,6720) *[4C]*	9477 (7720,14296) *[5M]*	7016 (5993,8265) *[5C]*	15535 (11243, 21963) *[1C]*	10979 (5012,31890)	5438 (4890, 6118)

Disappearances, $n = 867$							
	QS models		QS2/QS3 models		two-way model	Sample coverage approach	
	Marginal	Conditional	Marginal	Conditional	Conditional	HSC	LSC
2 NGOs, 5 govt	1391 (1267,1565) *[2M]*	1479 (1333,1661) *[2C]*	1382 (1249,1552) *[3M]*	1508 (1350,1706) *[3C]*	1940 (1413, 2976) *[1C]*	1164 (1052, 1342)	1079 (1019, 1162)
Collapsed NGOs, 5 govt	1397 (1256,1580) *[4M]*	1497 (1351,1696) *[4C]*	1397 (1258,1579) *[5M]*	1523 (1370,1735) *[5C]*	1975 (1428, 3054) *[1C]*	1267 (1126, 1483)	1133 (1064, 1227)

3. Population Size Estimation with Inactive Lists: Hierarchical mixture models and Missing Data with Application to Armed Conflict Data

[1]Shira Mitchell, [1,3,4]Al Ozonoff, [5]Kristian Lum, [2]Alan M. Zaslavsky, and [1]Brent A. Coull

[1]Department of Biostatistics, Harvard School of Public Health
[2]Department of Health Care Policy, Harvard Medical School
[3]Clinical Research Center, Boston Childrens Hospital
[4]Department of Pediatrics, Harvard Medical School
[5]Network Dynamics and Simulation Science Laboratory, Virginia Tech

Abstract

Since 1964, tens of thousands of people have died in Colombia's armed conflict. Underreporting the level of violence obscures the true nature of the conflict, precluding development of effective solutions. We develop hierarchical log-linear capture-recapture models to estimate the number of armed conflict killings that occurred in Casanare, Colombia in the years 1998-2007. Lack of data the early years motivates the use of hierarchical models that borrow strength across time. We investigate two methods to handle groups actively collecting data in different but overlapping time-periods. One fills in the inactive periods, treating the counts in those years as missing data. Another does not, instead incorporating the inactivity into the model. We compare these, as well as hierarchical versus unpooled models. A simulation study shows that the Bayesian hierarchical models have shorter confidence interval width, with similar or better coverage than the unpooled models, and show robustness to the exchangeability assumption. They enable us to obtain useful intervals for the number of killings in the early years, where there are less data, so we can look at trends across time that guide political analysis of the conflict. We provide guidance for capture-recapture studies with inactive lists and recommend the use of hierarchical modeling in capture-recapture.

3.1 Introduction

Since 1964, the Colombian armed conflict between the military, guerrilla, and paramilitary groups has killed tens of thousands of people, and displaced millions. Underreporting the level of violence obscures the true nature of the conflict, precluding development of effective solutions. Violence hidden from official reports and the press endangers the peace process by failing to hold perpetrators and policy-makers accountable.

Both government and nongovernment groups (NGOs) in Colombia report killings and disappearances. If we assume that a documented case from any list truly happened, then

no single list from the government or NGO is complete. In this paper, we use the statistical technique of *capture-recapture* to estimate the number of killings in the Casanare region of Colombia in each of the years 1998 to 2007, using data provided by six groups. Casanare is in the central eastern region of Colombia with a population of 350,000. It contains oil fields and a British Petroleum pipeline. Injection of cash into the economy from oil profits without government capacity for managing order created an environment conducive to violence by guerrilla and paramilitary groups (Davy et al., 1999). Human rights groups and policy-makers ask: How many killings occurred in Casanare? What are the trends across time?

With its basis in ecology at the turn of the twentieth century, capture-recapture (also known as multiple systems estimation) has also been used to estimate totals for human populations, see Fienberg (1992); for Disease Monitoring and Forecasting. (1995); Chao et al. (2001). Early work using capture-recapture for human rights data was done by HRDAG, the Human Rights Data Analysis Group. HRDAG used capture-recapture to estimate the number of killings and disappearances in Casanare (Lum et al., 2010; Guberek et al., 2010), but data sparsity made it difficult to provide estimates for the first two years, when there is little data (see Figure 2a in Lum et al. (2010)). Mitchell et al. (2013) compared marginal and conditional models for estimating the total number of killings and disappearances across all years. In this paper, we estimate the number of killings in each year, using models that borrow strength across time in order to obtain estimates in years with little data. Gelman (2006) shows by cross-validation that multilevel (hierarchical) modeling gives more accurate predictions than no-pooling and complete-pooling regressions. Capture-recapture does not lend itself to cross-validation, so we instead rely on a simulation study.

Fienberg et al. (1999) and Chao et al. (2001) discuss various approaches to modeling capture-recapture data. These include log-linear models that account for dependencies among lists introduced by Fienberg (1972) (see Bishop et al. (1975); Fienberg (2000)), ecological models that model the probability of capturing animal i in list j (see Chao et al.

(2001)), Bayesian hierarchical modeling approaches (Roberts, 1967; Smith, 1991; Castledine, 1981; George and Robert, 1992; Madigan et al., 1995; Madigan and York, 1997), and the sample coverage approach of Chao and Tsay (1998); Tsay and Chao (2001); Chao et al. (2001). Fienberg et al. (1999) explore relationships among the first three classes of this list. We fit log-linear models that can be motivated by a Rasch latent variable formulation for partial quasi-symmetry described in Fienberg et al. (1999), where nongovernment and government lists have different catching-probability distributions.

One challenge of the Casanare data is that the groups collect data with different levels of intensity across years. When it is known that some lists are inactive in certain strata, defined by years or regions, methods developed by Zwane et al. (2004); van der Heijden et al. (2009); Sutherland et al. (2007) may be appropriate. These methods treat inactive lists as missing data and use the Expectation Maximization (EM) algorithm to fill in what the list would have captured in the year or region in which it was inactive. As with other applications of EM, the algorithms make a missing at random (MAR) assumption, which is likely valid when the missingness provides no information about the underlying process, such as when a list is not yet established in a particular year. In our paper, we develop a Bayesian hierarchical model based on the EM algorithm in Zwane et al. (2004), using ideas from Dominici (2000).

In the Casanare killings data we analyze in this paper, organizations may have a harder or easier time collecting data as the violence changes. It is not known when or why a list is inactive. In the raw data, some lists have a bimodal distribution of observed killings. In some years the group only captures a low number and other years many. We model the capture intensity of a group over time as a mixture distribution with component distributions for inactive periods and active periods.

In Section 3.2 we describe the motivating dataset of violence records in Casanare. In Section 3.3 we describe candidate models we will compare. In Section A.1.1 we discuss fitting algorithms. In Section 3.5 we discuss simulated data. Section 3.6 presents results

from simulations and the real data. We then make conclusions and discuss future work.

3.2 Motivating Dataset - Casanare

Our data are lists of killings provided by 15 groups, both government and NGOs. These lists are called *sources* or *captures* in the capture-recapture literature. After de-duplication of records, there are 2629 reported killings. The six longest lists for killings combined report 2619 killings, missing only ten. Three of the six lists are from NGOs. A majority (1871) of records appear in only one list. For information about the groups, the matching algorithm to connect observations across lists, and raw data descriptives, see A.2.2.

Many zero cells in a table cross-classifying lists causes large standard errors and unstable results when fitting models (see Agresti, 2002, p.394). To alleviate this sparsity and because the longest lists contain most of the observed records, we take only the top six lists.

Table 3.1: Number of records in lists in each year (records not unique across lists), lists ordered longest to shortest for killings. Below each organization's acronym it is indicated whether it is a government organization (govt) or nongovernment organization (NGO).

year	IMLM (govt)	PN0 (govt)	VP (govt)	CCJ (NGO)	CIN (NGO)	CCE (NGO)
1998	1	0	0	14	13	3
1999	2	0	0	6	8	2
2000	213	0	5	22	23	0
2001	262	0	2	21	12	0
2002	268	1	0	33	9	0
2003	348	274	2	12	11	0
2004	412	324	295	14	11	1
2005	210	155	138	8	13	16
2006	104	71	26	3	2	15
2007	54	0	33	27	36	35

The raw data reveal that some lists appear to be either active or inactive in the collection of killings records across time. For example, National Institute of Forensic Medicine Deaths (IMLM) appears to be inactive during the first two years, but active from 2000 onwards.

The National Police (PN0) appears active only from 2003 to 2006, and Human Rights Observatory of the Vice Presidency (VP) from 2004 onwards. However, the NGO groups (CCJ,CIN,CCE) appear to operate at a lower level through much of the years, where the Colombia-Europe-US Coordination (CCE) appears inactive until 2005. This appearance leads us to incorporate a bimodal distribution into the list main effects of our model. We want to allow for borrowing strength across time through hierarchical modeling, but the assumption that all years are exchangeable for a particular list appears doubtful due to the active and inactive periods.

3.3 Candidate models

We consider log-linear models for the Casanare data presented in Section 3.2. We compare different log-linear models, a standard method in the capture-recapture literature (Fienberg, 1972).

In the *unpooled, zeros from sampling model (U-ZS)*, we treat all zero counts as sampling zeros except for the number of people missing from all lists. In other words, we do not treat inactive lists as missing data, as in Zwane et al. (2004); van der Heijden et al. (2009); Sutherland et al. (2007). In the *hierarchical, zeros from sampling model (H-ZS)*, we fit a hierarchical log-linear model that enables us to borrow strength across years and explicitly model an active versus inactive list structure. We also fit the *unpooled, zeros from missing data model (U-ZM)*, where we treat inactive lists as missing data, following the methods in Zwane et al. (2004). Finally, we fit the *hierarchical, zeros from missing data model (H-ZM)*, where we treat inactive lists as missing data in a hierarchical log-linear model that borrows strength across years.

For each of these log-linear models, we specify separate main effects parameters for each list and year combination, allowing the recording probability for each list to vary freely over time. Data sparsity makes fitting all two-way interactions difficult. Instead, moti-

vated by HRDAG's suggestion that dependence is lowest between lists of different types. We restrict interactions between two NGOs to be equal, between two government lists to be equal, and between a government and NGO to be equal (Mitchell et al., 2013).

Let J be the number of lists, or sources, which cover overlapping strata of the population defined by $t = 1, ..., T$. For our applications, the strata variable is year. Let $N^{(t)}$ be the size of the closed population in stratum t, for example, the number of killings in Casanare in year $t = 1998$. Let $n_{\mathbf{k}}^{(t)}$ be the number of population units in stratum t with recording pattern $\mathbf{k}$, a string of 1's denoting recording in a list and 0's denoting non-recording, of length J. We assume a multinomial sampling plan, independent for years $t = 1, ..., T$:

$$\big(n_{\mathbf{o}}^{(t)}, ..., n_{\mathbf{k}}^{(t)}, ..., n_{\mathbf{1}}^{(t)}\big) | N^{(t)}, \boldsymbol{\lambda}_t, \boldsymbol{\omega} \sim Multi\big(N^{(t)}, (\pi_{\mathbf{o}}^{(t)}, ..., \pi_{\mathbf{k}}^{(t)}, ..., \pi_{\mathbf{1}}^{(t)})\big)$$

where $\boldsymbol{\pi}^{(t)} = (\pi_{\mathbf{o}}^{(t)}, ..., \pi_{\mathbf{k}}^{(t)}, ..., \pi_{\mathbf{1}}^{(t)})$ are the multinomial probabilities, and $\boldsymbol{\lambda}_t$ and $\boldsymbol{\omega}$ are log-linear model parameters. Our log-linear models are of the form

$$\begin{aligned} \log \pi_{\mathbf{k}}^{(t)} = \lambda_{0t} + \lambda_{1,t} k_1 + ... + \lambda_{J,t} k_J + \\ \sum_{j,j' \in NGOs} \omega_{NGO} k_j k_{j'} + \sum_{j,j' \in govt} \omega_{govt} k_j k_{j'} + \sum_{j \in NGOs, j' \in govt} \omega_{mix} k_j k_{j'}. \end{aligned} \tag{3.1}$$

The H-ZM model also specifies

$$\lambda_{j,t} \sim N(\mu_j, \tau^2), \tag{3.2}$$

and the H-ZS model specifies a mixture model for the main effects, independent

$$\lambda_{j,t} \mid \gamma_{j,t} \sim (1 - \gamma_{j,t}) N(\mu_{inactive}, \tau^2_{inactive}) + \gamma_{j,t} N(\mu_j, \tau^2), \tag{3.3}$$

where $\gamma_{j,t} = 1$ if list j is active in year t and $= 0$ otherwise. We fix parameters $\mu_{inactive} = -9$ and $\tau^2_{inactive} = 3$ and estimate μ_j and τ^2 from the data, with the parameters given hyper-priors $\mu_j \sim N(0, \sigma_\mu^2)$, $\tau^2 \sim IG(a, b)$, where we take $a = 0.01$ and $b = 0.01$, a commonly used prior. We assign priors $\gamma_{j,t} \sim Bern(1/2)$, and $\omega_{NGO}, \omega_{govt}, \omega_{mix} \sim N(0, \sigma_\omega^2)$ independent, where $\sigma_\mu^2 = \sigma_\omega^2 = 100,000$ is very large so that the priors are essentially flat. For the unknown totals, we use the *single observation unbiased prior* (SOUP) $\pi(N^{(t)}) \propto 1/N^{(t)}$, see Meng and Zaslavsky (2002) and Section 3.4.4.

For the H-ZS model, we use the indicators for list activity in the interactions so that, for example, the interpretation of ω_{govt} is the log-odds ratio between two government lists for years in which both lists are active, conditional on the other lists. We have

$$\log \pi_{\mathbf{k}}^{(t)} = \lambda_{0t} + \lambda_{1,t}k_1 + ... + \lambda_{J,t}k_J + \sum_{j,j' \in NGOs} \omega_{NGO}\gamma_{j,t}\gamma_{j',t}k_j k_{j'} + \sum_{j,j' \in govt} \omega_{govt}\gamma_{j,t}\gamma_{j',t}k_j k_{j'} + \sum_{j \in NGOs, j' \in govt} \omega_{mix}\gamma_{j,t}\gamma_{j',t}k_j k_{j'}. \tag{3.4}$$

For U-ZS, a main effect for list j in year t is estimated to be a low negative number. For H-ZS, this main effect is estimated to come from the inactive portion of mixture 3.3, which is centered at a low negative number $\mu_{inactive}$. The H-ZS model makes the assumption that list main effects for inactive years are exchangeable and main effects for a given list for active years are exchangeable (Gelman et al., 2003). We also investigate an *AR1 hierarchical, zeros from sampling model (AR1-ZS)* where

$$\begin{bmatrix} \lambda_{j,1} \\ \vdots \\ \lambda_{j,T} \end{bmatrix} \mid \begin{bmatrix} \gamma_{j,1} \\ \vdots \\ \gamma_{j,T} \end{bmatrix}, \mu_j, \rho, \tau^2 \sim N \left(\begin{bmatrix} (1-\gamma_{j,1})\mu_{inactive} + \gamma_{j,1}\mu_j \\ \vdots \\ (1-\gamma_{j,T})\mu_{inactive} + \gamma_{j,T}\mu_j \end{bmatrix}, \right.$$
$$\left. \begin{bmatrix} \gamma_{j,1}\gamma_{j,1} & \gamma_{j,1}\gamma_{j,2}\rho & \dots & \gamma_{j,1}\gamma_{j,T}\rho^{T-1} \\ \gamma_{j,2}\gamma_{j,1}\rho & \ddots & \dots & \dots \\ \vdots & & & \\ \gamma_{j,T}\gamma_{j,1}\rho^{T-1} & & & \gamma_{j,T}\gamma_{j,T} \end{bmatrix} \tau^2 + \begin{bmatrix} 1-\gamma_{j,1} & 0 & \dots & 0 \\ 0 & & & \\ \vdots & & \ddots & \\ 0 & & & 1-\gamma_{j,T} \end{bmatrix} \tau^2_{inactive} \right). \tag{3.5}$$

The model specifies that when a list is active in years t and $t+r$ the main effects have correlation ρ^r, but if one is inactive and the other is active, or if both are inactive, their correlation is zero. We assign the same priors as for the hierarchical mixture model, with prior $\rho \sim Unif(0,1)$ for the correlation parameter.

In addition to the joint models across years, we also fit log-linear models to each year separately,

$$\log \pi_{\mathbf{k}} = \lambda_0 + \lambda_1 k_1 + ... + \lambda_J k_J + \sum_{j,j' \in NGOs} \omega_{NGO} k_j k_{j'} + \sum_{j,j' \in govt} \omega_{govt} k_j k_{j'} + \sum_{j \in NGOs, j' \in govt} \omega_{mix} k_j k_{j'}. \tag{3.6}$$

3.4 Fitting models

3.4.1 EM algorithm from Zwane et al. (2004)

To fit the U-ZM model, where we consider inactive lists as missing data, we implement the algorithm proposed by Zwane et al. (2004). We describe the algorithm via a general example to avoid the partitions notation in Zwane et al. (2004). If in year t, lists 3 and 4 are known to be inactive, we treat cell counts such as $n^{(t)}_{01000}$ as a margin $n^{(t)}_{01{+}{+}0}$, and we treat the four cells $n^{(t)}_{01000}$, $n^{(t)}_{01010}$, $n^{(t)}_{01100}$, $n^{(t)}_{01110}$ as missing data.

For each iteration i of the EM algorithm, for the E step we set

$$\hat{n}^{(t,i+1)}_{01010} = \frac{\sum_{p=1}^{T} \pi^{(p,i)}_{01010}}{\sum_{p=1}^{T} \pi^{(p,i)}_{01000} + \pi^{(p,i)}_{01010} + \pi^{(p,i)}_{01100} + \pi^{(p,i)}_{01110}} n^{(t)}_{01{+}{+}0}. \tag{3.7}$$

The standard EM algorithm requires a different E step (Dempster et al., 1977):

$$\begin{aligned}\hat{n}^{(t,i+1)}_{01010} &= E[n_{01010,t} | n^{(t)}_{01{+}{+}0}, \boldsymbol{\pi}^{(t,i)}] \\ &\text{which by properties of the multinomial is} \\ &= \frac{\pi^{(t,i)}_{01010}}{\pi^{(t,i)}_{01000} + \pi^{(t,i)}_{01010} + \pi^{(t,i)}_{01100} + \pi^{(t,i)}_{01110}} n^{(t)}_{01{+}{+}0}. \end{aligned} \tag{3.8}$$

Thus, the assumption being made by Zwane et al. (2004) is that the probability of a recording pattern (such as 01010) given that you are in the margin (such as 01++0) is identical across years. This is a stronger assumption than 'missing at random' (MAR), the usual assumption required for the EM algorithm.

For the M step of the algorithm, we fit the log-linear model to the completed data $\{n^{(t)}_{\mathbf{k}}\}_{\mathbf{k} \neq 00000, 00010, 00100, 00110}$.

3.4.2 Bootstrap Confidence Intervals

For calculation of confidence intervals of the population size for U-ZS and U-ZM models as well as the separate models in each year 3.6, we use the parametric bootstrap (Buckland

and Garthwaite, 1991; Norris and Pollock, 1996). We take 500 bootstrap samples from the multinomial distribution, using the maximum likelihood estimates of $\{\widehat{\boldsymbol{\pi}}^{(t)}\}_t$ from the data. We then refit the model to get estimates $\widehat{N}^{(b,t)}$ for all years t. With estimates from $b = 1, ..., 500$ we then take quantiles to get 95% intervals for each year t.

3.4.3 Fitting the hierarchical models

Computation for the hierarchical models is done by a Gibbs sampling algorithm. If we are fitting the H-ZS model, we sample from the posterior

$$\{N^{(t)}\}_t, \{\lambda_{j,t}\}_{j,t}, \boldsymbol{\omega}, \{\gamma_{j,t}\}_{j,t}, \{\mu_j\}_j, \tau^2, \underbrace{\rho}_{\text{for AR1}} \mid \{n_{\mathbf{k}}^{(t)}\}_{\mathbf{k}\neq\mathbf{o},t}, \tag{3.9}$$

where $\boldsymbol{\omega} = \{\omega_{NGO}, \omega_{govt}, \omega_{mix}\}$. If we are fitting the AR1 model, then the posterior includes the parameter ρ. The steps of the algorithm for the mixture and AR1 hierarchical models are given in Appendices A.2.1 and A.2.1.

If we are fitting the H-ZM model, we sample from the posterior

$$\{N^{(t)}\}_t, \{n_{\mathbf{k}}^{(t)}\}_{\mathbf{k}\neq\mathbf{o},t}, \{\lambda_{j,t}\}_{j,t}, \boldsymbol{\omega}, \{\gamma_{j,t}\}_{j,t}, \{\mu_j\}_j, \tau^2, \underbrace{\rho}_{\text{for AR1}} \mid \{n_{\mathbf{k}}^{(t)}\}_{obs}, \tag{3.10}$$

where $\{n_{\mathbf{k}}^{(t)}\}_{obs}$ are observed cells and margins when we view zero counts for lists in a year as missing data.

From the MCMC samples, we obtain estimates from the posterior mean and 95% posterior intervals from the 2.5% and 97.5% empirical quantiles of the MCMC realizations.

3.4.4 Single Observation Unbiased Prior

The Jeffreys prior $\pi(N) \propto 1/N$ used by Smith (1991); Castledine (1981) is a *single observation unbiased prior* (SOUP) for the population total N (Jeffreys, 1961; Meng and Zaslavsky, 2002). In other words, it has the property that the posterior mean of N conditional on

the observed data is unbiased. As shown by Stuart and Zaslavsky (2005), for the simple two-list case, the SOUP is uninformative for the list inclusion probabilities. We prove in Appendix A.2.1 that this extends to our log-linear model parameters where we take the yearly population total priors to be independent $\pi(N_t) \propto 1/N_t$.

3.5 Simulated Data

In the spirit of calibrated Bayesian analysis advocated in Rubin (1984), we check which of our log-linear models provides the least biased estimates with shortest 95% confidence or Bayesian posterior intervals with 95% coverage. The real datasets do not allow us to assess bias or coverage, so we turn to simulated data.

First, we generate data from the H-ZS model. We use the posterior means of the model parameters and population totals $N^{(t)}$ from fitting the Casanare data. When fitting U-ZM and H-ZM models to these data, we would treat cell counts such as $n^{(2003)}_{010100}$ as being $n^{(2003)}_{01+10+}$ if lists 3 and 6 are inactive in 2003.

We also generate data from the H-ZM model. In other words, we generate data from model 3.4 with prior 3.2, where all lists are always active. We then obscure the inactive lists, following the pattern of the Casanare data, taking sums for lists that are inactive so that only margins such as $n^{(2003)}_{01+10+}$ are visible if lists 3 and 6 are inactive in 2003. When fitting U-ZS and H-ZS models, we treat margins such as $n^{(2003)}_{01+10+}$ as cell counts $n^{(2003)}_{010100}$ if lists 3 and 6 are inactive in 2003. However, for the generated data we see that $\log(\pi^{(2003)}_{01+10+})$ is not linear in the parameters,

$$\begin{aligned}\log \underbrace{\pi^{(2003)}_{01+10+}}_{\text{treat as } \pi^{(2003)}_{010100}} &= \log\left(\pi^{(2003)}_{010100} + \pi^{(2003)}_{010101} + \pi^{(2003)}_{011100} + \pi^{(2003)}_{011101}\right) \\ &= \log\left(\exp(\lambda_{0t} + \lambda_{2,t}) + \exp(\lambda_{0t} + \lambda_{2,t} + \lambda_{4,t} + \lambda_{2,4}) + ...\right).\end{aligned}$$

Thus, our 0's from sampling models may have a disadvantage to the 0's from missing data models, which would have the correct model form.

We check robustness to the assumption of exchangeability, 3.2 for H-ZM and 3.3 for H-ZS, by generating data according to the AR1-ZS model.

3.6 Results

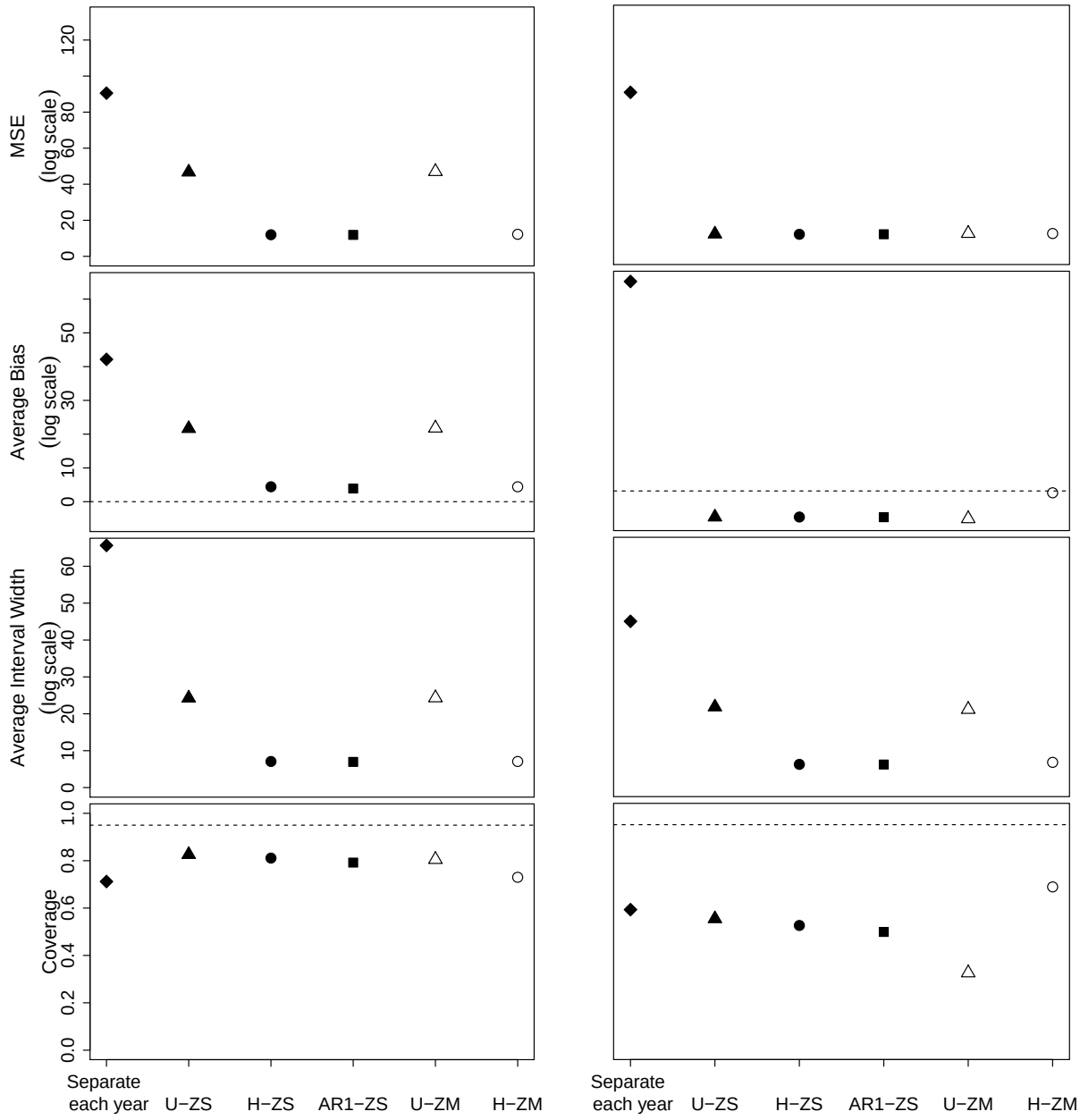
3.6.1 Simulation Results

We do 100 simulations from each of the H-ZS, AR1-ZS and H-ZM models. When simulating data from the H-ZS or AR1-ZS models (see Figure 3.1a and Figure A.17a in A.2.4), the hierarchical models H-ZS, AR1-ZS and H- ZM have smaller mean square error (MSE), bias, and interval width, achieving similar coverage to the unpooled models. We plot average bias as the $\log(\text{abs}(bias)) * \text{sign}(bias)$, enabling us to use the log-scale with negative values. Results are very similar whether data are generated by the AR1-ZS model with $\rho = 0.5$ versus the H-ZS model (equivalent to AR1-ZS with $\rho = 0$). Thus, there is robustness to the exchangeability assumption made by the H-ZS and H-ZM models.

When simulating data from the H-ZM model (see Figure 3.1b), all five joint models perform similarly in MSE and bias, though the H-ZM model is slightly less biased. Again, the hierarchical models have shorter interval widths. For coverage, none of the models achieve close to 95% coverage, but the H-ZM model does best. Note in particular that the U-ZM model, using methods from Zwane et al. (2004) has only 40% coverage. We suspect that the reason is related to the fact that their algorithm uses the E step 3.7 rather than 3.8. In extra simulations not included here, the U-ZM estimates have worse bias relative to U-ZS when fit to data from the H-ZS model with a large difference between capture intensity in active versus inactive periods.

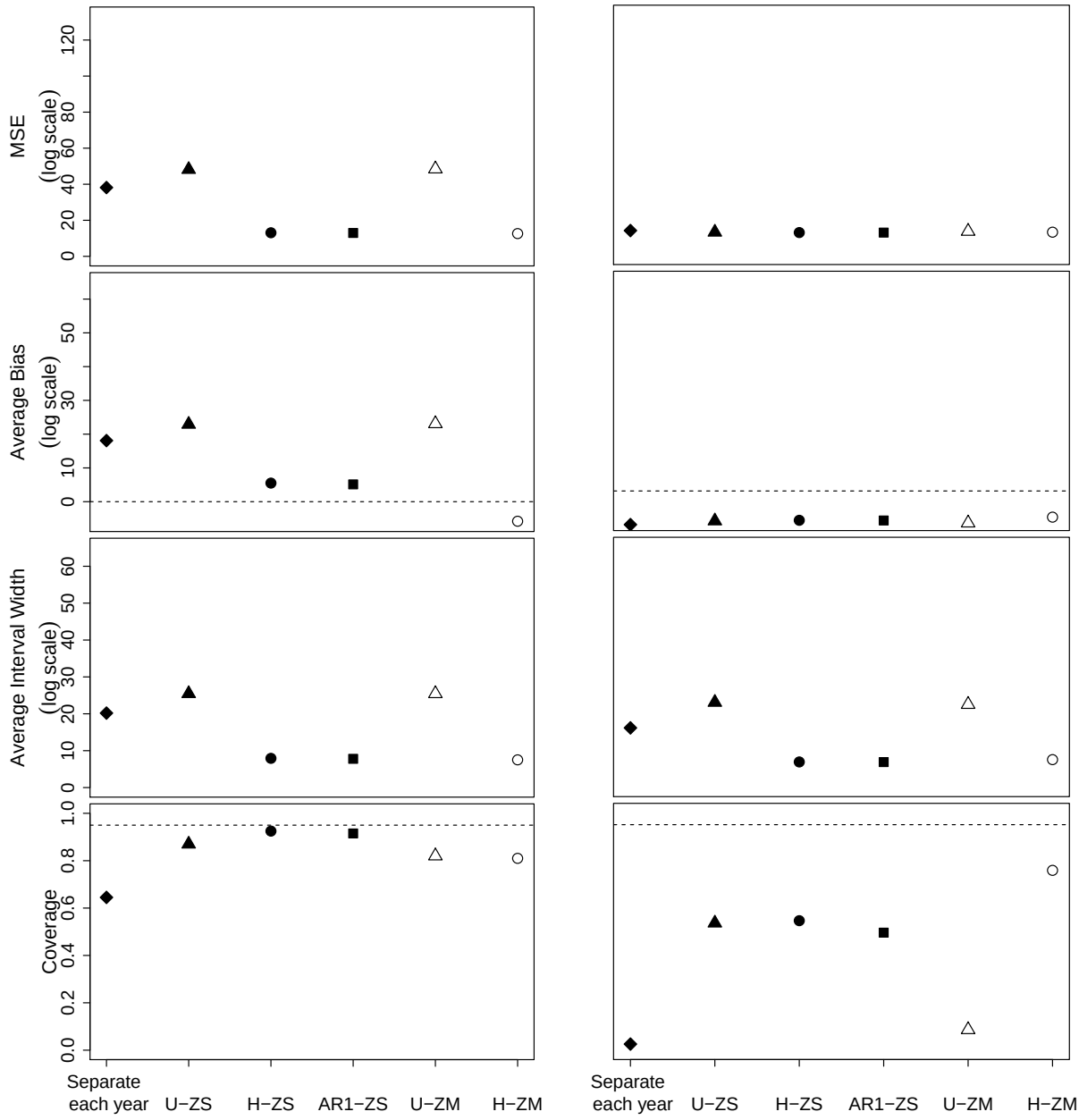
We also examine simulation results restricted to 1998 and 1999, when there is less data (see Figure 3.2 and in A.2.4, Figure A.17b). Simulating from the H-ZS model, we see that the H-ZM gives lower and negatively biased estimates in the first two years, and the coverage is close to the nominal 95% for the H-ZS and AR1-ZS models.

In A.2.4 we also look at how often the estimates $\widehat{N}^{(t)}$ are higher than the population of Casanare. When generating from the H-ZS or AR1-ZS models and fitting separate models in each year, this occurs 17% to 69% of the time in years 1998-2002, and rarely in later years. For data generated by the H-ZM model, this occurs 41% of the time in 2002, and rarely in other years. When fitting the unpooled models U-ZS and U-ZM, it occurs up to 25% of time during years 1998-2002 with data generated by H-ZS or AR1-ZS, and never in later years or when data is generated by the H-ZM model. Estimates for the hierarchical models H-ZS, AR1-ZS and H-ZM never have estimates exceeding the population of Casanare.



(a) H-ZS (b) H-ZM

Figure 3.1: Results from simulations, generating data from the H-ZS and H-ZM models, using μ_j, τ^2, ω, $N^{(t)}$, and $\gamma_{j,t}$ from posterior means of Casanare data. We do 100 simulations from each of the H-ZS and H-ZM models.



(a) H-ZS (b) H-ZM

Figure 3.2: Results from years 1998 and 1999 in the simulations, generating data from the H-ZS and H-ZM models, using μ_j, τ^2, ω, $N^{(t)}$, and $\gamma_{j,t}$ from posterior means of Casanare data. We do 100 simulations from each of the H-ZS and H-ZM models.

3.6.2 Real Data Results

We fit the joint models U-ZS, H-ZS, AR1-ZS, U-ZM, and H-ZM, as well as the separate models in each year, to the Casanare data described in Section 3.2. When fitting U-ZM and H-ZM, the methods of Zwane et al. (2004) require it be known when lists are missing. For the Casanare data, this information is not available, and we assume a list is missing if it has exactly zero counts in a year.

We see in Figure 3.3 that for the Casanare data, the hierarchical models give much shorter intervals than the other methods in 1998 and 1999, when few lists operate. The H-ZS model gives posterior mean $\widehat{\tau}^2 = 0.5$, H-ZM gives $\widehat{\tau}^2 = 2.6$, the AR1-ZS model gives $\widehat{\tau}^2 = 0.4$, and the correlation parameter $\hat{\rho} = 0.3$.

Our H-ZS model gives $\widehat{N} = 8999\ (7388, 11291)$ and AR1-ZS gives $\widehat{N} = 8275\ (6340, 10498)$ for the total killings in Casanare between 1998 and 2007, similar to the best-fitting model collapsing years from Mitchell et al. (2013), which gives $\widehat{N} = 9277\ (7538, 11720)$. The ZM models give lower estimates than the ZS models, with H-ZM giving $\widehat{N} = 7273\ (6077, 9044)$.

3.6.3 Posterior Predictive Checks

For the Casanare data, we want to check that the hierarchical models are consistent with the data. We do posterior predictive checks for the H-ZS, AR1-ZS and H-ZM models in Figure 3.4. We draw 500 simulated values from the posterior predictive distribution of replicated data, and compare these samples to the observed data through three test statistics. The first statistic is the number of events captured by only one list. The second is the number captured by exactly two lists, and the third is the maximum number of captures for any event. Note that with $J = 6$ lists, the highest this statistic can be is 6. We report Bayesian p-values, the probability that the replicated data could be more extreme than the observed data, where the probability is over the posterior distribution of the model

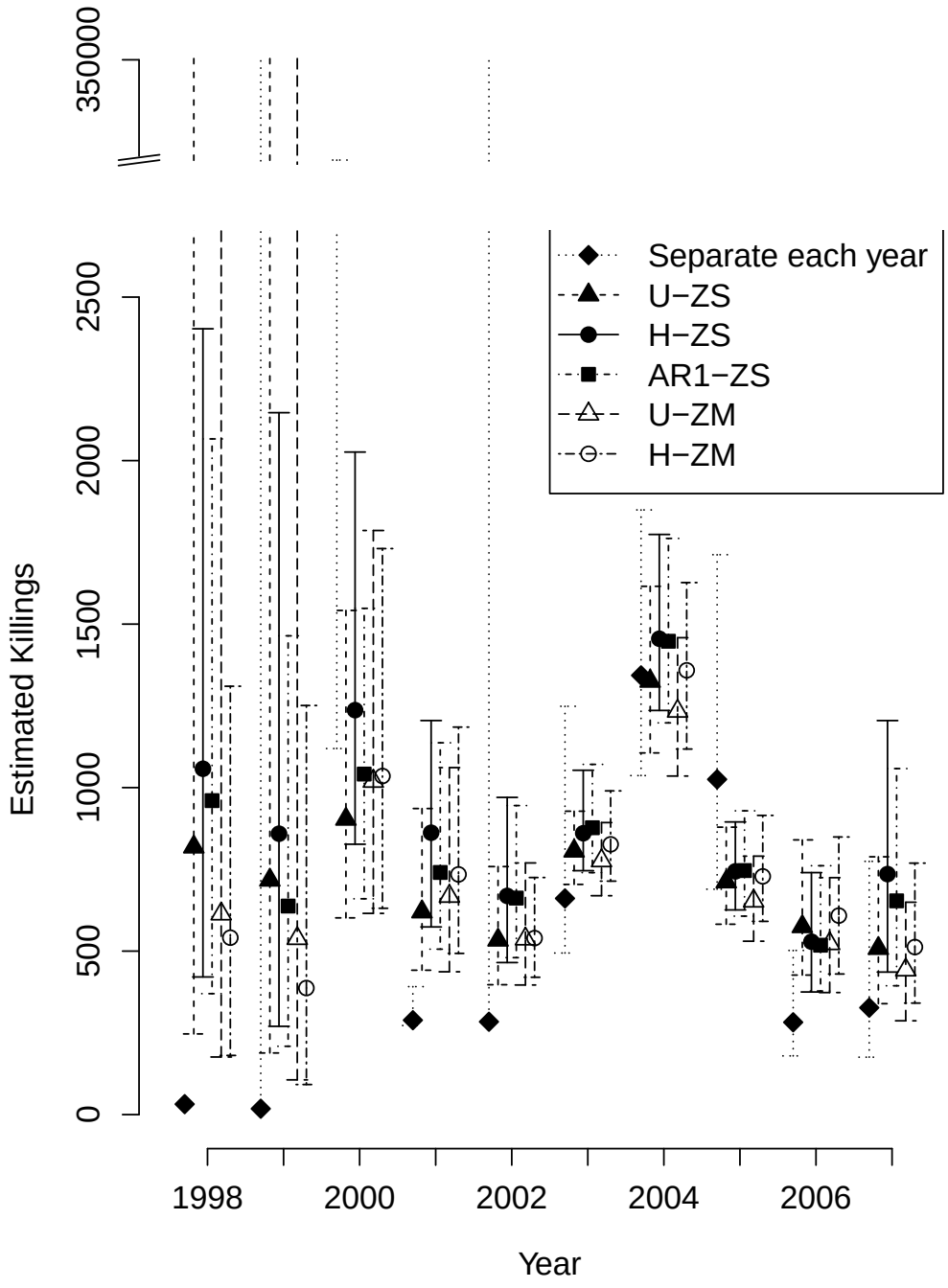


Figure 3.3: Point estimates and 95% intervals for the number of killings in Casanare, Colombia across 10 years, estimated using $J = 6$ lists. For U-ZS, $\widehat{N} = 7520\ (6309, -^*)$ for H-ZS, $\widehat{N} = 8999\ (7388, 11291)$ for AR1-ZS, $\widehat{N} = 8275\ (6340, 10498)$ for U-ZM, $\widehat{N} = 7005\ (5678, -)$ for H-ZM, $\widehat{N} = 7273\ (6077, 9044)$. *We use the symbol "$-$" to indicate that the upper limit is greater than the population of Casanare.

parameters and replicated data (Gelman et al., 2003, p.162). All p-values were not significant at the 0.05 level, indicating that the observed data looks plausible under the posterior predictive distributions from all three models. Graphical posterior predictive checks are available in A.2.3 and show good fit to the data, with visual assessment showing better fit for the H-ZS and AR1-ZS models.

3.6.4 Comparing the Casanare Data and Simulation Study

When simulating from the H-ZM model, the average estimates for the H-ZM model are higher than for the H-ZS and AR1-ZS models. Simulating from the H-ZS or AR1-ZS models, the three hierarchical models give similar average estimates, though in Figure 3.2a we see that the H-ZM gives lower and negatively biased estimates in the first two years. When fit to the Casanare data, H-ZM gives lower estimates than H-ZS and AR1-ZS in all years except for 2006. Thus, the behavior of the models when fit to the Casanare data more closely mimics our simulations from H-ZS and AR1-ZS in terms of average estimates. In all our simulations and the real Casanare data, the hierarchical models give shorter intervals, though the difference is mainly in the first two years.

3.7 Discussion

In this paper we investigate models to estimate the number of armed conflict killings that occurred in Casanare, Colombia in the years 1998-2007. Earlier work estimated the total number of killings across the ten-year period (Mitchell et al., 2013) or was unable to provide an estimate for the early years due to lack of data (Lum et al., 2010; Guberek et al., 2010).

The methods proposed in Zwane et al. (2004) require knowledge of when lists are active or inactive, and make the missing at random assumption and an assumption that the probability of a recording pattern given that you are in a margin is identical across

years. We fit a hierarchical, zeros from missing data model that makes the first of these assumptions and borrows information across the years. When applying these methods to the Casanare data, we made the ad hoc assumption that a list is inactive only if it has a count of exactly zero in a given year. Since the Casanare data include years with very low activity, only a handful of events recorded, it may not be compelling to treat these years differently than the zero counts. The hierarchical, zeros from sampling model and the autoregressive extension allow for the active and inactive periods to be estimated from the data, grouping very low activity together with zero counts. These models have an advantage over the standard log-linear models in interpretation, because the interaction terms are log-odds ratios between lists when both are on. Furthermore, these models allow for borrowing of information across years, enabling us to obtain useful intervals for the number of killings in the early years, where there are less data.

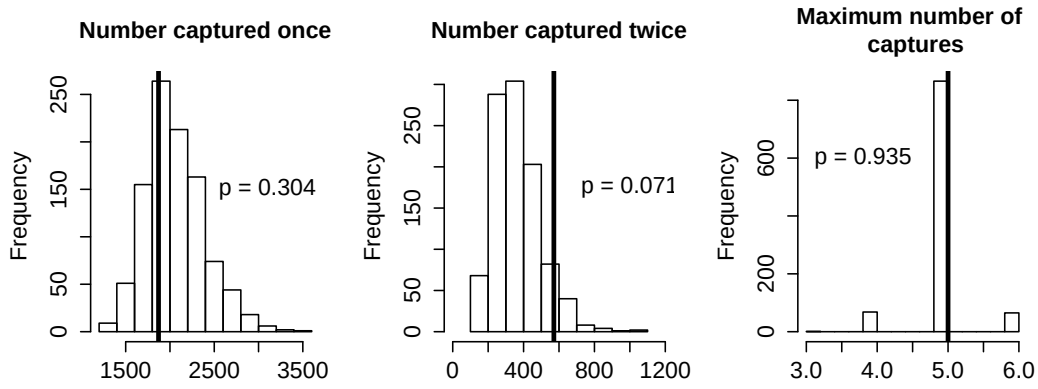
Posterior predictive checks give reason to trust the estimates from all the hierarchical models, whether we treat the view the zeros as from sampling or missing data. As noted in Section 3.6.4, the behavior of the models when fit to the Casanare data more closely mimics our simulations from the hierarchical, zeros from sampling models in terms of average estimates. In these simulations, the hierarchical, zeros from sampling models give better coverage than the zeros from missing data models.

For the Casanare data, we prefer the hierarchical, zeros from sampling models both because it is a more principled approach to handling the years with low activity, and because our simulation study gives us reason to trust these models more for these data. This leads us to believe that the level of violence in 1998 and 1999 was higher, and more similar to the level in 2000, before lowering in 2001. The question is then, why did the violence go down in 2001? The lower estimates for 1998 and 1999 from the zeros from missing data models would instead of lead us to ask why the violence increased temporarily in 2000.

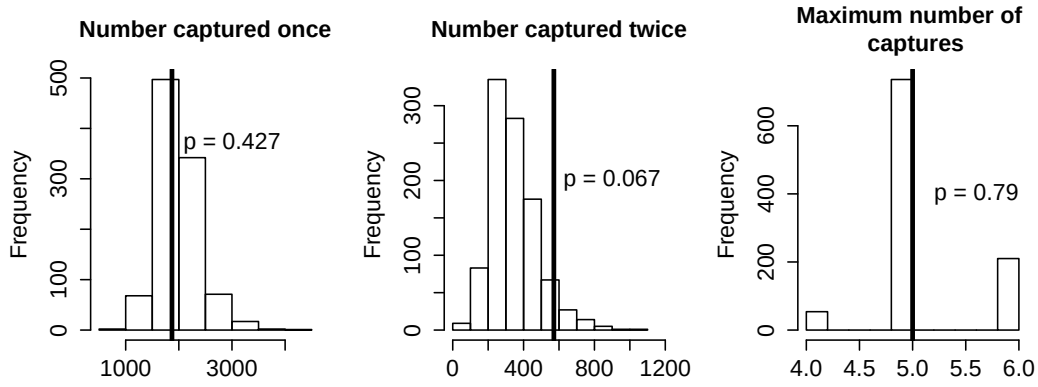
In epidemiology and human rights applications, it is common for lists to concentrate their efforts in different years, locations, or segments of the population. If these times,

locations, or groups are overlapping, then the methods in this paper can be useful. If there is a temporal or spatial structure, we recommend using a hierarchical model to borrow strength across years and regions in a principled way. Autoregressive (temporal or spatial) should be explored and assessed via posterior predictive checks.

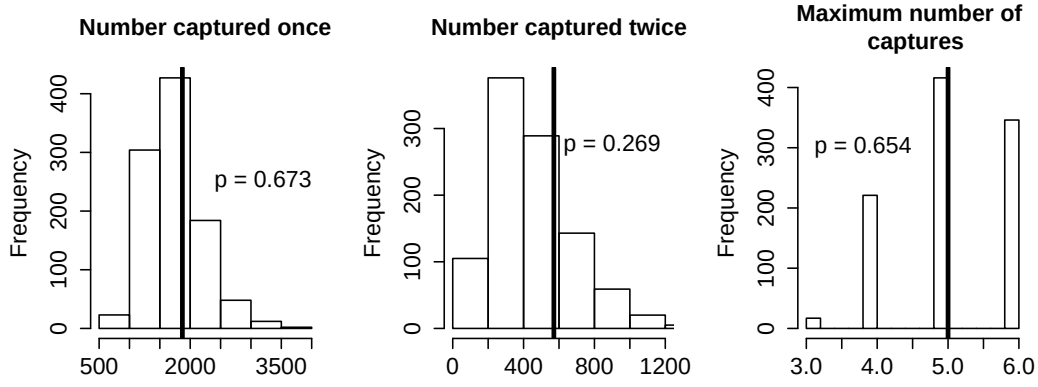
The hierarchical Bayes formulation of the Rasch model in Fienberg et al. (1999) can be extended to estimate population totals over strata, such as years. Their Bayesian latent variable formulation of the partial quasi-symmetry models discussed by Darroch et al. (1993), can be used to model the different visibility of events in NGOs versus government lists. Our log-linear models instead focus on the list dependence, modeling the way in which NGO and government groups may be sharing information.



(a) H-ZS model posterior predictive checks.



(b) AR1-ZS model posterior predictive checks.



(c) H-ZM model posterior predictive checks.

Figure 3.4: Posterior predictive distribution, observed result, and p-value of three test statistics for the Casanare data, fitting the hierarchical models. Based on 500 simulations. The first statistic is the number of events captured by only one list. The second is the number captured by exactly two lists. The third is the maximum number of captures for any event (with $J = 6$ lists, the highest this can be is 6).

4. The Millennium Villages Project: A protocol for the final evaluation

[1]Shira Mitchell, [2,3]Andrew Gelman, [4]Uyen Kim Huynh, [4]Maria Muniz, [4]Xiaoyi An, [4]Lucy McClellan, [5]Alan M. Zaslavsky, [6]Joseph Blitzstein, [4]Paul Veldman, [4]Eva Quintana, [4]Cheryl Palm, [4]Elizabeth Katwan, [4]Saira Qureshi, [4]Andrew Thorne-Lyman, [4]Sonia Ehrlich Sachs, [4]Jeffrey D Sachs

[1]Department of Biostatistics, Harvard School of Public Health
[2]Department of Statistics, Columbia University, New York
[3]Department of Political Science, Columbia University, New York
[4]The Earth Institute, Columbia University, New York
[5]Department of Health Care Policy, Harvard Medical School
[6]Department of Statistics, Harvard

Abstract

This document is the protocol for the end-line evaluation of the Millennium Villages Project. In Section 4.1 we provide some of the project's background. In Section 4.2 we describe the project's principles and site selection, in Section 4.3 we outline the core evaluation questions and in Section 4.4 we outline the sections that will address them: the adequacy assessment in Section 4.8, impact evaluation in Section 4.9, cost assessment in Section 4.10, and process evaluation in Section 4.11. In Section 4.5 we defined the primary outcomes, and in Section 4.6 the secondary outcomes. In Section 4.7 we describe the survey data collection. In Section 4.12 we discuss our plan for transparency. Section 4.13 outlines the timeline for the evaluation. Section 4.14 addresses ethical issues, and Section 4.15 addresses study limitations. Taken together, this evaluation protocol is designed to assess the MVPs model of integrated rural development on achieving the MDGs in 10 selected sub-Saharan African study sites.

4.1 Background

Rural sub-Saharan Africa is home to millions of the world's poorest people, most of whom live in below-subsistence conditions and face acute social and economic vulnerabilities and a high burden of disease (Kifle et al., 2002). Deficiencies in food, education, and income are compounded by limited access to adequate housing, water and sanitation, transport, and communication services. Taken together, these act both to increase exposure and to reduce resistance to disease and avoidable death (Sachs et al., 2004).

In September 2000, world leaders came together at the UN Millennium Summit to adopt the Millennium Declaration, committing their nations to a new global partnership to reduce extreme poverty, setting targets with a deadline of 2015 that have become known as the Millennium Development Goals (MDGs) (United Nations, 2000).

The UN Millennium Project, an independent advisory effort from 2002-2005 initiated by UN Secretary-General Kofi Annan, identified steps designed to achieve the MDGs (Sachs and McArthur, 2005). The project recommended investments in scientifically-driven interventions, in the context of open, well-governed, and market-based economies.

The Millennium Villages Project (MVP) was initiated in 2005 to help design, measure, and scale up effective delivery systems for the UN Millennium Project's recommended interventions across multiple sectors (Sanchez et al., 2007). The MVP was piloted in Sauri, Kenya and Koraro, Ethiopia in 2005, and expanded to include fourteen villages with half a million inhabitants by 2006. The project's goal was to help rural populations achieve the MDGs and move the village towards self-sustaining economic growth.

4.2 Project description

The MVP model for achieving MDGs in rural, sub-Saharan Africa adheres to several core principles:

- The implementation of multi-sectoral and integrated interventions grounded in well-managed delivery systems;
- The implementation of scientifically-driven technologies and practices;
- The participation of local village communities in the planning, execution, and monitoring of a set of interventions, designed specifically for each Millennium Village (MV);
- Co-planning and implementing the MV concept at the local and district level government agencies;
- Cost-sharing with government, donors, and the community;
- Learning by doing in the design of intervention systems.

The MVP multi-sector approach includes interventions in food production, nutrition, education, health services, roads, energy, communications, water supply and sanitation, enterprise diversification, environmental management and business development. See Appendix A.4 for a listing and timeline of the core MVP interventions by sector. The MVP delivers diverse, simultaneous interventions to address multiple objectives and to enable possible synergistic gains through positively interacting interventions, motivated by the idea that the whole may be greater than the sum of its parts (Sachs, 2007; Sachs et al., 2004). We discuss evaluation of the intervention synergies in Section 4.9.8. The MVP uses technologies and techniques such as agroforestry, insecticide-treated malaria bednets, antiretroviral drugs, community deworming, remote sensing, and geographic information systems.

The MVP is a ten-year project with two five-year phases. The first phase concentrates on "quick win" interventions, which include:

- Free mass distribution of malaria bednets and effective antimalarial medications;
- Elimination of user fees for primary schools and essential health services;
- Expansion of school meals programs; and
- A large-scale replenishment of soil nutrients to smallholder farmers on lands with nutrient-depleted soils.

At the end of this initial phase, roughly in 2009 (with variation across the sites) the MVP evolved from a demonstration of quick wins to focus more on commercializing the gains in agriculture and on designing local service delivery systems in health, education, infrastructure, agriculture and business development.

4.2.1 MV Study Site Selection

The project grew organically at the start. It began as a single site, Sauri, Kenya and then expanded to 10 countries by 2007. By 2007, there were a total of 14 MV sites. The criteria for selecting these as MV sites were as follows:

- All sites were located in 'hunger hotspots', defined as areas with more than 20% of underweight children under the age of five (Sanchez et al., 2005).
- The MV sites were selected to represent ten different agroecological zones in sub-Saharan Africa, each with distinctive agronomic, health and economic challenges.
- The sites were selected with the agreement and commitment of national governments to partner with MVP's design and implementation of the projects model.

Of the 14 MV sites, 12 were selected to be research sites (See Figure 4.1). We define research sites as those where baseline and at least one follow-up household survey data was collected. The two MV sites not part of the 12 research sites are: Toya, Mali, and Gumuliira, Malawi. These two sites were not selected as part of the MV research sites due to two main reasons: 1) Two MV research sites (Tiby, Mali, and Mwandama, Malawi) were selected as already representing these particular agroecological zones, 2) Toya and Gumuliira have populations less than 10,000 inhabitants preventing them from becoming sites with reasonable economies of scale to operate the full MVP package of interventions.

Ikaram, Nigeria, and Dertu, Kenya, stopped operating interventions because of financing shortages and were only funded for the first phase of the MVP project period. Since then, the Ikaram MV has been under the administration of the local Nigerian government. Dertu, Kenya, faced civil war in addition to financing problems. There is no longitudinal data being collected from these four sites. One can consider our estimates of program impact as estimates for the superpopulation of villages in which MVP treatment would not have been disrupted by financing shortages or political instability.

Ikaram, Nigeria, and Dertu, Kenya, stopped operating interventions because of financing shortages and were only funded for the first phase of the MVP project period. Since then, the Ikaram MV site has been under the administration of the local Nigerian government. The Dertu MV site operations were disrupted by civil war in 2010, in addition to financing problems. All four of these sites have not carried out any data collection since 2008. There are no plans to collect household survey from these four sites when the project finishes its operations at Year 10. Maybe we should say something about TOT (treatment on treated) versus ITT (intention to treat) here? Or just this: One can consider our estimates of program impact as estimates for the super population of villages in which MVP treatment would not have been disrupted by financing shortages or political instability.

We restrict this evaluation to include the ten research sites still in operation: 1) Pampaida, Nigeria, 2) Tiby, Mali, 3) Potou, Senegal, 4) Bonsaaso, Ghana, 5) Sauri, Kenya, 6) Ruhiira, Uganda, 7) Mayange, Rwanda, 8) Mbola, Tanzania, 9) Koraro, Ethiopia, and 10) Mwandama, Malawi. See Table 4.1 for the project start dates of each research site.

In these ten clusters, there is an average population of approximately 45,000 inhabitants per cluster. In each cluster, interventions commenced in an area of approximately 5000 inhabitants (roughly 1000 households). See Table 4.1 for a more detailed description of the ten clusters. This area, referred to as the *research village*, or the *MV1*, receives the most MVP interventions. Over time, as additional resources became available, the MVP expanded to surrounding villages, referred to as the *MV2*, which receives a subset of the interventions implemented in MV1. The MV1 and MV2 combined constitute the MV cluster.

Table 4.1: Description of the ten MV's covered in this evaluation. These population counts were collected in 2010-2012 via a census in MV1, and household counts in MV2 (2010-2013).

MV	Agroecological Zone	Start date	DHS dates	Number of Households in MV1 (research village)	Population in MV1 (research village)	Number of Households in MV1+MV2 (cluster)	Population in MV1+MV2 (cluster)
Koraro, Ethiopia (EK)	Highland Mixed	Q1 2005	2000, 2005, 2011	1171	5914	16,620	67,711
Bonsasso, Ghana (GB)	Tree Crop	Q3 2006	2003, 2008	1201	6049	5555	25,257
Sauri, Kenya (KS)	Maize Mixed (bimodal)	Q1 2005	2003, 2008-9	996	5112	13,685	67,315
Mwandama, Malawi (MM)	Cereal-Root (Southern miombo)	Q3 2006	2000, 2004, 2010	889	3598	9038	37,153
Tiby, Mali (MT)	Cereal-Root (Sudan savanna)	Q3 2006	2001, 2006	986	14,290	5529	80,131
Pampaida, Nigeria (NP)	Agro-silvopastoral	Q2 2006	2003, 2008	924	6244	4152	28,057
Mayange, Rwanda (RM)	Root Crop (miombo)	Q3 2006	2000, 2005, 2010	726	3343	5724	25,710
Potou, Senegal (SP)	Agro-silvopastoral	Q2 2006	2005, 2010-11	717	7227	3137	32,823
Mbola, Tanzania (TM)	Maize Mixed (unimodal)	Q2 2006	2004-5, 2010	1041	6952	5972	37,024
Ruhiira, Uganda (UR)	Highland Perennial	Q2 2006	2000-1, 2006, 2011	1159	5663	9948	46,570

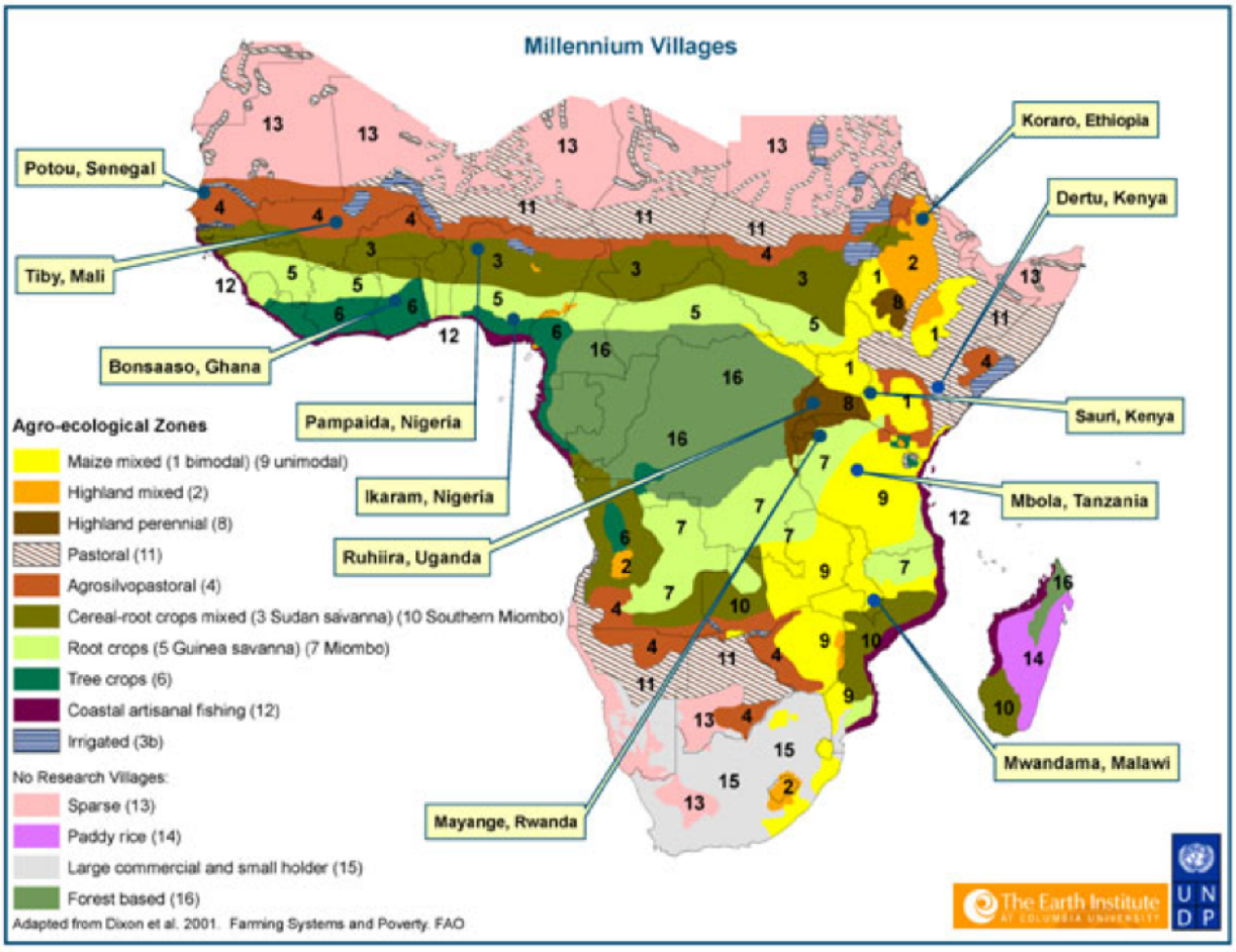


Figure 4.1: Millennium Village Project study sites.

4.3 Evaluation Questions

While specific interventions within the MVP package have scientifically proven effect (Kremer and Holla; Cohen and Dupas, 2010; Schofield, 2014), the specific package of interventions and its implementation elicit many questions of interest:

1. Are the **MDG targets** met within each research village (MV1) site?
2. What are the MVP **treatment effects** on each of the primary MDG indicators of interest? In other words, what progress towards the MDGs is attributable to the program?
3. Does the MVP stay within the **target of $120** annual per capita cost?
4. Which factors have most barred or best facilitated the implementation of integrated intervention packages? What are the biggest **lessons learned**?

Our evaluation aims to answer these questions using mixed methods outlined below. Section 4.15 discusses other big questions related to the MVP that will not be answerable by this evaluation.

4.4 Project Evaluation Components

The final evaluation uses mixed methods to answer the questions in Section 4.3.

1. **Adequacy assessment:** To assess the adequacy of reaching MDGs in the research villages (MV1).
2. **Impact evaluation:** To attempt to isolate the effect of the program in the research villages (MV1). In other words, to answer the question of causality.

3. **Cost assessment:** To compute all annual on-site costs of carrying out MVP interventions and activities in each of the sites - by sector, year, stakeholder, and MV1 versus MV2. It will assess the costs relative to the projects $120 annual per capita cost-sharing model.

4. **Process evaluation:** To document and assess the design and implementation of the multi-sector MVP approach, generating new insights regarding project feasibility by documenting the content of interventions, their timing and sequence, and key barriers and facilitators to their introduction - providing lessons and highlighting challenges for maintenance of the MVP delivery systems and interventions, and transfer to other contexts.

We elaborate on each of these components in Sections 4.8, 4.9, 4.10, and 4.11, respectively.

4.5 MDG Primary Outcomes

The primary outcomes of interest, for both the adequacy assessment and impact evaluation (see Sections 4.8 and 4.9), are a subset of fifteen MDG indicators. These primary outcomes are listed in Table 4.2, with definitions from standard UN MDG guidelines and 2015 targets.

Excluded indicators are provided in Appendix A.6. These include indicators inapplicable in the context of the villages (e.g. proportion of seats held by women in national parliament, proportion of urban population living in slums or the official development assistance and global market access indicators); indicators that are too difficult or costly to measure (e.g. CO2 emissions, total, consumption of ozone-depleting substances, or HIV prevalence among population aged 15-24 years); indicators that are not part of the core MVP interventions (e.g. literacy rates of 15-25, since MVP education related interventions focus on primary aged children). Finally, some indicators are excluded because

there are insufficient sample sizes to capture the indicator (e.g. maternal mortality), yet are tracked by the project.

To facilitate comparability, the MDGs are assessed using survey tools that draw directly from international assessment tools for program monitoring and evaluation including the USAID funded Demographic and Health Surveys (DHS), UNICEFs Multiple Indicator Clusters Surveys (MICS) and World Bank Living Standards Measurement Study (LSMS) survey.

Table 4.2: The Millennium Development Goal (MDG) indicators that constitute our primary outcomes. All indicators are defined in MDG (2014). Targets are defined by the UNDP (Group, 2003), unless otherwise indicated: $^{(m)}$ denotes a target defined by the MVP (MVP, 2009), for indicators without specific 2015 targets set by the UNDP; $^{(u)}$ indicates a target defined by UNESCO (UNESCO Institute for Statistics, 2010). A * on the indicator number labels those indicators measured by DHS. $^{(f)}$: The Foster-Greer-Thorbecke, or FGT, metric is a generalized measure of poverty within an economy. The general formula is $FGT_\alpha = \frac{1}{n}\sum_{i=1}^{q}\left(\frac{z-y_i}{z}\right)^\alpha$. We note that indicator 1.1 is FGT_0.

#	Indicator	2015 Target	Definition, village-level	Variable in the Causal model
MDG Goal 1: Eradicate extreme poverty and hunger				
1.1	Proportion of population below 1.25 USD (PPP 2005) per day	Reduce to 50% of the level in 1990	**proportion** of all people that live below 1.25 USD (PPP 2005) per day	**average** USD (PPP 2005) per day people live on. At the individual-level, outcome is **continuous**.
1.2	Poverty Gap ratio	Reduce to 50% of the level in 1990	$FGT_1^{(f)} = \frac{1}{n}\sum_{i=1}^{q}\frac{z-y_i}{z}$ summing over all q people below the poverty line, $z = 1.25$ USD (PPP 2005), where y_i is income of person i, and n is the number of people sampled	use the above ⇑

Continued on next page

Table 4.2 – continued from previous page

#	Indicator	2015 Target	Definition, village-level	Variable in the Causal model
1.8^*	Underweight among children under 5 years old	Reduce to 50% of the level in 1990	**proportion** of children under 5 years old who fall below minus two standard deviations of weight for age of the WHO standard	**average** weight for age z-score among children under 5. At the individual-level, outcome is **continuous**.
MDG Goal 2: Achieve universal primary education				
2.1^*	Net attendance ratio in primary education	$\geq 90\%^{(m)}$	**proportion** of children of primary school age who attend primary or higher education	$\Leftarrow$ same. At the individual-level, outcome is **binary**.
2.2	Proportion of pupils starting grade 1 who reach last grade of primary education	$\geq 90\%^{(m)}$	**estimated probability** of a student in grade 1 advancing to the end of primary school, subject to retention rates in the year of the survey, estimated by the *reconstructed cohort method* (MDG, 2014)	$\Leftarrow$ same. No individual-level outcomes.
MDG Goal 3: Promote gender equality and empower women				
3.1^*	Gender parity in primary education	$0.97 - 1.03^{(u)}$	**ratio** of girl gross attendance ratio to boy gross attendance ratio [gross attendance ratio = (# of people in primary school)/(# of children of primary school age), note that this can be greater than 1]	$\Leftarrow$ same. No individual-level outcomes.
MDG Goal 4: Reduce child mortality				

Continued on next page

Table 4.2 – continued from previous page

#	Indicator	2015 Target	Definition, village-level	Variable in the Causal model
4.1*	Under-5 mortality rate	Reduce to 33% of the level in 1990	**estimated probability** of a child dying before age 5 years subject to survival rates in the 5 year window preceding the survey (assumption: constant survival rate in that 5 year window); usually reported as deaths per 1000 live births	$\Leftarrow$ same. Survival analysis at individual-level, using birth histories, see Appendix A.7.3.
4.2*	Infant mortality rate	Reduce to 33% of the level in 1990	**estimated probability** of a child dying before age 1 year subject to survival rates in the 1 year window preceding the survey (assumption: constant survival rate in that 1 year window); usually reported as deaths per 1000 live births	$\Leftarrow$ same. Survival analysis at individual-level, using birth histories, see Appendix A.7.3.
4.3*	Measles immunization rate of 1 year-old children	$\geq 90\%^{(m)}$	**proportion** of children aged 12-23 months who received measles vaccine before their first birthday	$\Leftarrow$ same. At individual-level, outcome is **binary**.
5.2*	Skilled birth attendance	$\geq 70\%^{(m)}$	**proportion** of women age 15-49 years with a live birth in the last 2 years who were attended by a skilled health personnel during their most recent live birth	$\Leftarrow$ same. At individual-level, outcome is **binary**.

Continued on next page

Table 4.2 – continued from previous page

#	Indicator	2015 Target	Definition, village-level	Variable in the Causal model
5.3*	Modern contraception use	an absolute increase of 25% from the MVP Baseline per site$^{(m)}$	**proportion** of women age 15-49 years who are currently married or in a union where she or her partner is using a modern contraceptive method	$\Leftarrow$ same. At individual-level, outcome is **binary**.
5.5*	Antenatal care coverage - at least four (4) visits with any provider	$\geq 70\%^{(m)}$	**proportion** of women age 15-49 years with a live birth in the last 2 years who received antenatal care at least four times during their last pregnancy (with any provider)	**average** number of antenatal care visits per woman. At individual-level, outcome is a **count**.
MDG Goal 6: Combat HIV/AIDS, malaria and other diseases				
6.7*	Children under 5 sleeping under insecticide-bed nets	$\geq 80\%^{(m)}$	**proportion** of children under 5 years old who slept under an insecticide treated mosquito net the night prior to the survey	$\Leftarrow$ same. At individual-level, outcome is **binary**.
MDG Goal 7: Ensure environmental sustainability				
7.8*	Access to improved drinking water	Reduce proportion without access to 50% of the level without access in 1990	**proportion** of all persons who use an improved source of drinking water	$\Leftarrow$ same. At individual-level, outcome is **binary**.
7.9*	Access to Improved sanitation	Reduce proportion without access to 50% of the level without access in 1990	**proportion** of all persons who use improved sanitation facilities	$\Leftarrow$ same. At individual-level, outcome is **binary**.

4.6 Secondary Outcomes

While the 15 primary outcomes are MDG indicators, some non-MDG indicators may be of interest, including:

- fertility rate,
- assets (Filmer and Pritchett, 2001; Michelson et al., 2013),
- agricultural output,
- stress and mental health,
- occupation types,
- stunting,
- ownership (as opposed to use of) bed nets, and
- malaria mortality.

4.7 Survey data collection

Data for analyses are derived from population-based surveys at multiple points in time, routine monitoring systems, qualitative data, and economic cost data. Surveys are collected from the MV1 research villages. Qualitative process data and costing data are collected at the cluster-level (MV1 and MV2 as well). Here we describe survey data collection since baseline (2005). These survey methods will be mirrored within comparison villages at end-line (2015). For sample size calculations for the end-line survey, see Section 4.9.6. Economic costing data and qualitative process data are described in Sections 4.10 and 4.11, respectively.

4.7.1 Household surveys

Household selection: Within each MV1, a detailed household mapping was conducted at baseline (2005-2006), prior to the initiation of interventions. This process included a household and population census, Global Positioning System (GPS) readings, and household wealth ranking. Following this process, proportional sampling was used to randomly select 300 geographic and wealth-stratified households within the MV1 to undergo detailed periodic assessments.

Consenting households are followed longitudinally over two assessment rounds (Year 3, Year 5, and Year 10). Before each survey, a census is conducted. In the event of refusals or household attrition, a replacement household, present at baseline and from a similar baseline wealth strata, is chosen at random to maintain the sample size. In addition, households not present at baseline are added so that the fraction of new households in the sample equals the fraction of new households in the MV1 population at the time of the survey. This way, we can estimate progress towards the MDGs either among households present at baseline, or among a changing group of households that are present in the village. We note that the age distribution of households present at baseline may be skewed towards older people, so the cross section at end-line may be more relevant.

The Household Survey is administered to all household heads of sampled households (and/or other knowledgeable household member) capturing information on household demography, education, employment, malaria bed net usage, land ownership and use, agricultural and non-agricultural sources of income, assets, expenditure, consumption and access to basic services including water and sanitation, and energy, transport and communication.

In and Out Migration

At each survey round, the project does a census of MV1, and asks household (HH) heads whether each member from a previous round was:

- still living in HH,
- deceased,
- moved to another HH within the village,
- moved to another HH outside of the village,
- left HH to go to school outside of the village,
- child born to HH Member while member living in HH,
- moved into this HH from within the village, or
- moved into this HH from outside the village.

With the questions above, someone who leaves and returns at a later survey round would get counted twice, once in the out-migration and once in the in-migration. In the 2015 surveys, we propose to include questions to ask each surveyed individual if they have lived in either the cluster, or specifically MV1, since baseline. These questions combined with the above should allow us to compute statistics on in and out migration from the clusters, or more specifically, from the MV1 research sites.

For the impact and adequacy assessments, we propose to analyze those present since birth or baseline (whichever came first) separately from the cross-section of those present in 2015. This is likely more relevant that the separate analyses of houses present since baseline or not.

We will compute 2010 and 2015 statistics for:

- % of people who left MV1 since 2005,
- % who left the cluster since 2005,
- % of residents currently in the cluster who moved to the cluster since 2005,
- % of residents of MV1 who moved to MV1 since 2005.

4.7.2 Adult surveys

Within each participating household, all household members age 15-49 are given the adult survey. Household members are defined as those who have lived in the household for at least 3 of the past 12 months, and who 'normally eat from the same pot.' Additionally, the main provider for the household and newlyweds are given the adult survey as well, if they are age 15-49.

The survey examines health-related MDGs, nutrition and food security, and health seeking behavior.

4.7.3 Reproduction and pregnancy surveys

Birth histories are collected for all women in MV1 in each survey round, in order to estimate under five and infant mortality .

4.7.4 Biological and Anthropometric data

Biological testing: Tests for malaria (thick and thin smears) and anemia using a HemoCue point-of-care device (HemoCue Worldwide, 2014) is conducted among all children under 5 years of age, within sampled households.

Anthropometric data: Weight, height and mid-upper arm circumference is assessed using standard protocols among all children under 5 years of age, within sampled house-

holds.

Quantitative data collection and management

Enumerators have been hired and trained prior to previous survey rounds. For the final evaluation, two weeks of refresher training will be conducted. Surveys will be administered after an informed consent process and will be administered verbally. All questionnaires will be checked for quality three times post-enumeration and sent back to the field as needed. Random household visits will be undertaken by field supervisors to ensure quality control.

Data entry uses a template developed in CSPro (US Census Bureau, 2013) containing a series of pre-programmed range, skip, and logic checks to minimize errors in data capture. Double data entry will be undertaken for key indicators to reduce errors in data capture. Data cleaning will be conducted concurrent to data entry, using CSPros batch edit functionality that allows an additional series of data checks to be performed. Basic tabulation of MDG indicators will take place using CSPro, with data exported to Stata (StataCorp, 2011) employed for more complex analyses.

4.8 Adequacy Assessment

Adequacy evaluations assess how well a program met the expected objectives (Habicht et al., 1999). They require no control groups, and only depend on comparison of with previously established adequacy criteria. The MDGs serve as the established adequacy criteria for the MVP. From its inception, the project's standard was "adequacy" in achieving the MDGs, rather than optimality (McArthur et al., 2011). The standard was decided upon because of the project's standing as one of the only (if not the only) projects attempting to achieve all MDGs in a large and varied rural sub-Saharan African setting. Therefore, an adequacy assessment will be conducted in each of the MVP villages, to measure

progress towards pre-determined MDG targets established by the project. Measurement of the MDG outcomes will take place in 2015, at the end of the project, following 10 years of intervention exposure.

The extent to which progress towards the MDGs is attributable to the project will be assessed in the next section.

4.8.1 Targets

See Table 4.2 for the explicit numerical targets and by whom they were established. Seven targets are defined by the UNDP (Group, 2003), seven targets defined by MVP (MVP, 2009), and one by UNESCO (UNESCO Institute for Statistics, 2010). Seven targets were set relative to country-specific national rural 1990 baselines. Seven were defined as absolute targets. One target (for modern contraception use) is defined relative to the MVP baseline per site.

For the country baselines, we use the national rural averages for 1990, or a data point closest to the year 1990, whenever 1990 data are not available, see Table A.7 in Appendix A.5. If there is no separate breakdown of rural and urban, we simply use the national average. Reference data were compiled from a variety of sources, including the World Bank, World Health Organization, the Demographic Health Surveys (DHS), and United Nations Statistics Division databases, see Table A.7. See Table A.8 for the village-specific targets.

4.8.2 Sample sizes

For sample size recommendations we use the power calculations from the impact evaluation, see Section 4.9.6 below.

4.8.3 Data analysis

For each research village (i.e. MV1), point estimates and 95% confidence intervals will be computed and reported for each indicator, using data collected in 2015. These intervals will be compared against the targets defined in Table A.7.

4.9 Impact Evaluation

In this section we describe the impact evaluation for the MVP. By *impact evaluation*, we mean a measurement of the program's effect with great attention to determining true causal relationships. An impact evaluation attempts to prove that the stated "effect", "result", "impact", or "achievement" of a program represents the difference between what happened with the program and what would have happened without that program (Clemens and Demombynes, 2011). The MVP was not designed as a controlled evaluation. This controversial decision was justified on the basis of a focus on adequacy and feasibility, and logistical and ethical complexities including how to present the study to communities not receiving the interventions. Rigorous impact evaluation was overlooked while launching the complex, multi-country intervention.

But there has been a continuing call for such an evaluation (Clemens and Demombynes, 2011; Clemens et al., 2012; Butler, 2012; Nature editorial, 2012; The Economist, 2012; Starobin, 2013; Clemens and Demombynes, 2013). We believe an evaluation will be useful not only for assessing "statistical significance" of each treatment effect, but for estimating the *magnitude* of those effects. We hope to learn which indicators are most affected by the program. Additionally, methodology developed to overcome the challenges of this evaluation may inform the design of future evaluations.

In this section we discuss the challenges of such an evaluation, the design we have chosen, and the range of findings we anticipate. We first discuss the history of the project and mid-term reports. Then we outline the design, and in subsequent sections go into the

data sources in candidate comparison areas, the matching procedure, candidate causal models, power calculations, and identification of externalities and treatment synergies. Appendix A.7 goes into the technical details of the proposed impact evaluation.

History of the project's design

The project is a village-level intervention, and could not be randomized to individuals. However, randomization to *villages* was theoretically possible (Clemens and Demombynes, 2011). The project was not assigned to a random subset of candidate villages, producing the possibility that the MVs differ in observed and unobserved characteristics from other potential villages.

The project's gradual expansion from one village to ten was uncertain at the start, as it was unclear how much funding would be available. Randomizing one village to treatment and one to control does not create treatment and control groups similar enough to provide reliable causal inference without additional assumptions.

With ten treatment villages, as was the eventual size of the project, a design better equipped to identify a causal effect would have been: at baseline, select groups of areas within each country that match as closely as possible on geographic and poverty characteristics (possibly using some baseline surveys to assess comparability of areas), and randomly assign treatment to one area per country. This would have freed us from some untestable assumptions, and would have provided good baseline data in control areas. Operations of other NGOs and government programs could have been documented in the control areas. "Control" does not imply denial of any government programs, for example, bednet distribution. Rather, "control" refers to the fact that those areas do not (yet) get the full package of services provided by MVP. In this protocol, we use the terms "comparison" and "control" areas interchangeably.

We hope to produce the most credible analysis given the project's history, using expert-recommended statistical methods, and being clear about assumptions.

4.9.1 Mid-term Reports

Various articles about the evaluation of the MVP include MVP (2010), Clemens and Demombynes (2011), Pronyk et al. (2012), and Wanjala and Muradian (2013). Each leaves unanswered questions and critiques.

The project released its first public report in June 2010 (MVP, 2010). The report computed after-minus-before comparisons at the MVs. Despite the fact that the estimates were descriptive and not an attempt at impact evaluation, regrettably the report used the word "impact", which understandably caused confusion. The question of interest, what impact has the project had, is answered only under the assumption that the trend in outcomes in the absence of the intervention would have been flat. Though the report did stress that the results were preliminary, it did not state this assumption as a strong caveat.

Clemens and Demombynes (2011) contrasted the reported effects with estimates from a difference-in-differences analysis. They looked at three MVs, using rural households in the region where the MV is located as a comparison group, whose before and after data was obtained from the DHS. Clemens and Demombynes (2011) do not provide intervals of uncertainty for the difference-in-differences estimates, and the analysis adjusts for no covariates. The crucial assumption of *additivity* is needed with this strategy: in the absence of the MVP intervention, the differences in outcome over time would be the same across MVs and comparison areas (Gelman and Hill, 2007). This assumption can be made more believable by adjusting for covariates through matching and regression, see Appendix A.7.3.

Pronyk et al. (2012) also used difference-in-differences methods (for outcomes for which retrospective questions could provide baseline data), with adjustment for covariates via both matching and regression. However, there were concerns about the usefulness of the comparison villages, due to possible differences between comparison villages and MVs in political buy-in and the unclear selection procedure (Clemens and Demombynes, 2011;

Bump et al., 2012). This evaluation will make the selection of comparison areas rigorous and transparent, see below.

Wanjala and Muradian (2013) used a method related to ours (see Section 4.9.2), combining propensity score methods with regression estimation to look at the treatment effect in the Sauri, Kenya MV. They appear to adjust for variables that may be affected by treatment, which may be a source of bias, see Rosenbaum (1984). Their analysis assumes no village effects, attributing differences between the MV and comparison areas only to the treatment (we revisit this point in Appendix A.7.3).

These reports are all from the first half of the project, whereas the final evaluation will be the first to consider the project in its entire 10-year context.

4.9.2 Design

The project operates in ten village clusters in ten distinct countries in sub-Saharan Africa. The project has not systematically collected data in comparison areas. Comparison areas are key to defining and estimating the causal effect of the MVP. At end-line, in 2015, funding is available for surveying comparison areas. Each of the ten countries containing an MV is divided into districts (with local names used for the comparable term "district"), and each district contains several villages. Relevant comparison villages may be likely to be in the same district, as this will help control for local government, critical features of ecology, national markets, disease epidemiology, and other covariates.

In the absence of randomization to MVP treatment, establishing causal claims about the impact of the MVP relies on untestable assumptions. One of the most common assumptions in observational studies is that the distribution of potential outcomes (outcomes that would have happened for each village with MVP or without) be the same for MVs and comparison villages, once we control for confounding variables. This key assumption is known as *(strong) ignorability*, *unconfoundedness*, *no unmeasured confounders*, or *selection*

on observables (Rubin, 1976, 1978, 2008; Imbens and Rubin, 2014; Gelman and Hill, 2007; Greenland et al., 1999; Bang and Robins, 2005; Angrist and Pischke, 2009).

To make unconfoundedness as plausible as possible, we want to control for many variables which are not affected by treatment (Rosenbaum, 1984)). For our design, we follow *matching* with *regression*, following the advice of Rubin (1973); Rubin and Thomas (2000); Gelman and Hill (2007); Ho et al. (2007); Kreif et al. (2011); Abadie and Imbens (2011). The combination of the two methods is more robust than each alone (methods that use both treatment and outcome models are sometimes referred to as "doubly robust", see Robins et al. (2000); Robins and Rotnitzky (2001); Bang and Robins (2005); Imbens and Rubin (2014)). Matching serves to make the treatment and control groups more similar, with more overlap in covariates. This avoids using the regression to extrapolate to areas of poor overlap, which would rely heavily on the correctness of the linear model.

We want to select comparison villages that match, as closely as possible, the MVs at baseline. As mentioned above, funding limits us to surveying comparison villages within one district per country. To inform our selection, we need measures of variables in candidate comparison areas at baseline (and possibly post-treatment variables we are confident cannot be affected by MVP, such as rainfall). In Section 4.9.3 we discuss available data sources.

In Section 4.9.4 we propose to use propensity scores to select, for each MV, good comparison villages within the same country (Rosenbaum and Rubin, 1983b). After the matching procedure, we will have groups of treatment villages (the MVs) and matched comparison villages. We will then analyze the data using *multilevel regression*, adjusting for variables used in the matching.

Our regression models will either adjust for baseline outcome (often known as ANCOVA methods), or regress the difference in outcomes over time on other baseline covariates (difference-in-differences methods). We will compare these methods in Appendix A.7.3. If we include enough background variables to satisfy unconfoundedness, matching and

regression in combination should do well to approximate results from a randomized experiment (Dehejia and Wahba, 1999; Dehejia, 2005; Shadish et al., 2008).

Outcomes

Our primary outcomes are the indicators defined in Table 4.2, using the variables defined in the rightmost column of that table. In total, we have 14 primary outcomes of interest.

We alter some of the indicators, taking the raw form of the data rather than a dichotomized version, so as not to lose power (Rosyton et al., 2006; Gelman and Park, 2008). For the first two indicators, we prefer to use the raw data on income per day. For indicator 1.8, we prefer to use the weight for age z-score, rather than its dichotomized version. Lastly, for indicator 5.5 we take the number of antenatal care visits, rather than dichotomizing it. We will also report results from dichotomized versions using logistic regressions.

MV1 only

For the primary impact evaluation, we consider only the MV1, the core research village in each MV that receives the full set of interventions. Each MV1 contains roughly 1000 households.

4.9.3 Data in candidate comparison areas

We require baseline (2005) variables in the ten countries, measured at a fine enough geographic scale to be able to identify good comparison villages. Of particular importance are geographic data (agroecological zones, distance to a main road, distance to a town) and the primary outcomes of interest, 12 of which are measured by the Demographic and Health Surveys (DHS). Most of the indicators collected by MVP are also collected by the DHS, see Table 4.2, using similar survey tools (Muniz et al., 2011; Rutstein and Rojas,

2006).

Combining census and DHS data - small area estimation

DHS provides Global Positioning System (GPS) data for each surveyed cluster, usually a census enumeration area (EA). However, the disadvantage of DHS data is that it is geographically sparse, with roughly 350-900 EAs sampled out of 8000-600,000 EAs per country. Within each EA, about 20-40 households are sampled. On average, each EA has a population of 50-250 households. Thus, each MV1 is roughly the size of 4-20 EAs.

For a given country, if we have access to census data, we can combine census and DHS data in a *small area* model that estimates variables of interest at the EA or other small area level (Elbers et al., 2003; Rao, 2003). Balk et al. (2004, 2005) linked GPS clusters from DHS data in African countries to geographical databases in analyses of child mortality and malnutrition. Fujii (2005) combined data from the 2000 Cambodian DHS with the 1998 census to do *small area estimation* (SAE) to estimate the child malnutrition prevalence at the commune-level. Simler (2006) combined the 1991-92 Tanzanian DHS, 1988 census, and geographical variables to do small area estimation to estimate the height-for-age and weight-for-age z-scores at the district-level in Tanzania. Johnson et al. (2010) combined data from the 2003 Ghanaian DHS with the 2000 census to do small area estimation to estimate the proportion of institutional births at the district level. Mansour et al. (2012) explore spatial uncertainty in the DHS to census linkage. We propose to draw on the experience of these studies in our small area estimation procedures.

Beyond the primary outcomes of interest, we will identify additional variables collected by the DHS and censuses to use in the selection of comparison villages. For all these variables, for each of the ten countries, we will fit small area models (Ghosh and Rao, 1994; Ghosh and Natarajan, 1999; Nadram, 2000; Rao, 2003; Jiang and Lahiri, 2006). See Appendix A.7.1 for the small area models we propose and additional complications surrounding the use of small area estimates for selection of comparison areas.

Other surveys

Besides DHS, other survey data sources include UNICEFs Multiple Indicator Clusters Surveys (MICS) and World Bank Living Standards Measurement Study (LSMS) survey. In 2006, there is MICS data for Malawi and Ghana, and in 2007 there is data for Nigeria. In 2004-5 there is LSMS data for Malawi, and in 2004, LSMS data is available in the Kagera region of Tanzania (not the same region as the MV but may include some areas in the same agroecological zone).

Unlike DHS, MICS and LSMS do not report GPS coordinates of the sampled EAs, they only identify the district. Thus, they are not usable as survey data for small area estimation. However, it is possible to include district-level aggregates from these data sources as district-level covariates in models 4.1 and 4.2 in Appendix A.7.1.

Geographical data

In addition to survey data, we will collect, for the ten countries: geographical data including agroecological zones, soil type, rainfall, elevation, and distance to roads and towns from 2005.

4.9.4 Selecting comparison villages

If assignment to treatment is unconfounded given covariates, then assignment is unconfounded given the *propensity score*, the average assignment probabilities for subpopulations with a common value of the covariates (Rosenbaum and Rubin, 1983b). It is often simpler to find close matches using a scalar (the propensity score) rather than all covariates jointly. We want to match exactly on country and agroecological zone. Thus, for each MV, we will look at the estimated propensity scores for areas within the same country and agroecological zone. The areas in this group with estimated propensity scores close to the estimated propensity scores of the MV will be defined the "best matches".

First, we require estimates of propensity scores. We follow the conventional approach in the literature and use logistic regression on our baseline covariates (Imbens and Rubin, 2014). See Appendix A.7.2 for the proposed propensity score model.

In addition to these analyses, we will attempt to meet with the people who decided which villages would be MVs. Their reasoning will be reported alongside the propensity scores described above. We will also research which development programs operate in these candidate areas. These sources of information will guide the discussion of the choice of comparison areas, which we will open up to the scientific community, as we discuss in Section 4.12, below.

Without census data

If there are countries for which we are unable to obtain georeferenced census data, the propensity score model will be fit without estimates from small area models. Variables from geographical databases will still be included in the model. If census data exist for a subset of countries, we can do small area estimation in those countries for the outcome indicators measured by DHS. If we see good balance between MV and comparison area indicators from the countries with available census data, we can be more confident about the validity of our impact estimate, even though our candidate causal models discussed below would not be able to adjust for these indicators at baseline. This would hurt the efficiency of the procedures, increasing our posterior uncertainty. Thus, we will make acquiring georeferenced census data a priority.

4.9.5 Candidate Models for Causal Inference

We suggest a few types of causal models in Appendix A.7.3 that we propose to fit to the end-line outcome data. The analysis will fork in many ways, with different modeling choices. In the end-line evaluation we will report and compare all results to reduce the scope for fishing (i.e. deciding to report a model based on the realization of the conclu-

sion, see Humphreys et al. (2013)).

Causal inferences can be biased if we adjust for variables affected by treatment (Rosenbaum, 1984), so we restrict to adjusting for baseline variables or variables such as rainfall, which cannot be affected by the MVP. We have panel data in the MVs, but not in comparison areas. Due to anonymizing in the external (i.e. not MVP-conducted) surveys, we will not be able to identify the individuals who were surveyed at baseline in order to resurvey them at end-line. Therefore, we are limited to adjusting for aggregate baselines. See Appendix A.7.3 for the proposed models.

4.9.6 Power Calculations and Sample Size recommendations

See Appendix A.7.4 for power calculations, which were done using simulation methods.

For the mortality outcomes, all women in the research villages (MV1) and comparison villages will be surveyed to provide a birth history. For other outcomes measured by the household and adult surveys, we turn to the simulations in Figures A.19, A.20, and A.21 of Appendix A.7.4, while keeping in mind the simplifications outlined in A.7.4. Almost always, when more data are available, more can be learned, without any sharp cutoffs that point to a specific optimal choice.

4.9.7 Externalities

As outlined by ITAD (2013), there are three potential types of externalities. The first is a spread of services to nearby areas, reduction of infection risk, and externalities through local markets. Second, spending within the district containing the MV may shift from the MV to other areas in the district. Third, imitation of MVP interventions and adoption of policies such as bednet and fertilizer distribution.

To quantitatively estimate the first type of externality, we would want to sample within

the MV district at a walkable distance to the MV, but outside of it. If we sample close enough to the MV, we may be able to assume that the baseline would closely match that of the MV. In particular, though no baseline surveys were done in MV2, if we are willing to assume baseline was similar to baseline in MV1, we can estimate the treatment effect in MV2. This treatment effect would combine externalities from MV1 with the subset of interventions implemented in MV2.

The treatment effects estimated by the impact evaluation can be interpreted as the effect of the program *beyond* externalities such as policy changes. These effects inform whether the MVP package of interventions should be scaled up in settings where policies such as bednet and fertilizer distribution are already in place.

4.9.8 Estimating Treatment Synergies

In the absence of an experimental design including arms with all possible treatment combinations, it is very difficult to establish synergistic treatment interactions. ITAD (2013) propose to compare cost effectiveness for each component intervention of the MVP with cost effectiveness for similar interventions by other programs. A synergistic effect should imply that each component of the MVP is more effective per dollar than for the singular intervention of another project.

There are serious challenges with this approach. Differences in cost effectiveness could result from program design and implementation rather than synergies. The extent to which such comparisons can be made will also depend on how precisely we can estimate the MVP treatment effects for different outcomes, and how precisely the comparison programs can estimate their treatment effects. There will likely be too much noise to get a meaningful result.

We will not embark on this quantitative assessment of synergies.

4.9.9 Software

All multilevel models will be fit using the `lme4` package and Stan in `R`, (LME4 Authors, 2013; Stan Development Team, 2013; R Development Core Team, 2014).

4.10 Cost Assessment

A fundamental hypothesis of the project is that the MVP package of interventions can be delivered at a modest cost. The needs assessment conducted by the UN Millennium Project estimated that achieving the MDGs would require local service delivery and community-based investments of approximately $US 120 per person per year (in 2005 USD) during the 10-year period from 2005 to 2015 (Sanchez et al., 2005; World Health Organization, 2003).

It is important to understand that the $120 is not an increment above a baseline level of spending. Rather it is the estimated total cost of the MDG package of interventions, some part of which is in place without the MVP. The MVP is therefore providing a financial "top up" to existing funds, with the aim of reaching a total of around $120 per person per year. The incremental "cost" of the MVP, is therefore not the full $120, but only the top up. Will we not know how big the top up is relative to the full $120, unless we do costing in the comparison areas, which at the moment is not included in our evaluation budget. We do know that the roughly $60 per person per year that the project spends is part of the top up.

It is also important to underscore that this $120 annual per capita figure does not reflect the entire cost of the MVP project. The $120 includes the costs of service delivery, implementation, and on-site management, including estimated values of in-kind donations. Off-site costs, comprising salaries and overhead for all scientific and support staff at the Earth Institute and Millennium Promise staff based in New York and at the regional MDG Centers in Dakar and Nairobi, are excluded from this cost assessment. These excluded

off-site staff are primarily involved in project design, implementation research, monitoring and evaluation, logistics, and fundraising. They are not involved in direct operations, so their costs should be considered a one-time cost to design and operate the project, rather than an ongoing cost of running an MVP-style project in a scale-up context.

4.10.1 Methodology

Costs are collected for the entire project area (MV1 and MV2), but a distinction between MV1 and MV2 costs will be drawn, allowing for the costs in the research village (MV1), where investments have been more heavily concentrated, to be distinguished and analyzed separately. Due to the varying degree of detail in external stakeholders expenditure records, as well as the spillover effect of certain investments and the difficulty of isolating beneficiary groups, it will not be possible to distinguish perfectly between MV1 costs and MV2 costs for every intervention. In cases where estimations are necessary, all assumptions made will be recorded and clearly outlined in the final evaluation.

Of the estimated $120 per capita annual cost, the project (Millennium Promise) supplies $60 per capita, while national and local governments, external donors (including NGOs, multilateral organizations, and private donors), and the local community (mainly in kind as labor and material inputs) supply the rest (Figure 4.2) (UN Millennium Project, 2005). Understanding these inputs is critical to evaluating the success of the project in relation to the $120 per capita annual project target, and to assess scalability of project interventions (Sanchez et al., 2007). The nature and intensity of inputs is likely to differ substantially between clusters due to community needs, local disease profile, and local economic base.

A full economic costing assessment, in line with established methods of social and health policy interventions (Catterall, 1985; Ahren, 1976; Pushpangadan, 1997; Rahman and Alam, 1987; Hutchinson, 1969) is underway in each project cluster. The aim of the assessment is to document the annual on-site costs of the project by site, stakeholder (see Figure 4.2), sector (see Figure 4.3), year, and within the MV1 only as well as the entire

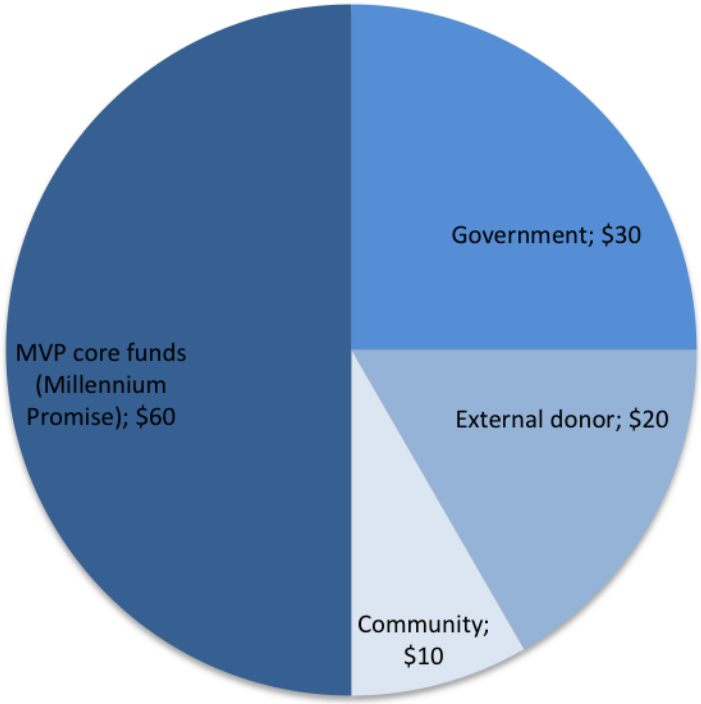


Figure 4.2: Costing model by stakeholder (2005 USD) (Sanchez et al., 2005).

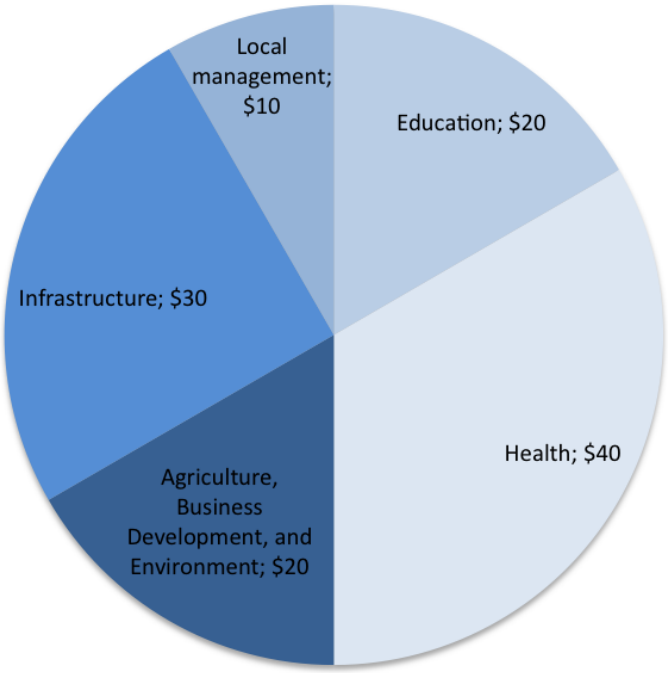


Figure 4.3: Costing model by sector (2005 USD) (Sanchez et al., 2005).

cluster.

Core project investments and external stakeholder investments are tracked via two complementary mechanisms. Core project expenditures (those made with funds that flow through the Millennium Promise bank account) are tracked and reported quarterly via the projects internal cost-tracking system. Expenditures made by external stakeholders (government, community, and other donors) within the sectors of education; health; agriculture, animal husbandry, business development, and environment; and infrastructure, are collected periodically by local site team members.

A series of data collection templates have been created for each stakeholder operating within each project cluster (including both MV1 and MV2). There are approximately 20 government, donor, and community stakeholders per project cluster. All costs within the defined project sectors (see above) are collected. The costing data collected via these individual stakeholder templates are amalgamated with the core project costing data collected via the internal tracking system to form a single comprehensive costing database for each project cluster.

For contributions made in kind, all prices are documented using the standard cost imputation method recommended for multi-center interventions (Grieve et al., 2009; Schulenburg, 2000; Wordsworth et al., 2005). This method involves establishing local unit costs for each in-kind contribution (e.g. daily wage rate in the case of labor contributions). These unit costs are then used along with qualitative data collected during key informant interviews to calculate a total cost for each contribution (e.g. daily wage rate $\times$ number of laborers $\times$ number of days worked) (Grieve et al., 2009; Schulenburg, 2000; Wordsworth et al., 2005).

4.10.2 Data management and analysis

After the costing data from all stakeholders have been collected and aggregated, the data will be archived and available for analysis. For the purposes of cross-site comparison, compatibility with core project expenditures, and measurement against the MVP costing model, all expenditure amounts will be converted to 2005 US dollars using average annual exchange rates for each project year. Annual per capita expenditures will be calculated for each project cluster, research village (MV1), sector, and stakeholder, using the total cluster population (MV1 and MV2) as well as the research village population (MV1 only).

Analysis of the data will be focused around questions of sustainability, replicability, and scalability. Per capita costs by sector and stakeholder are essential to planning any project scale-up or replication.

4.11 Process Evaluation

Program evaluations necessitate the inclusion of qualitative data in a mixed methods design to fully understand the effects of complex projects such as the MVP. Process Evaluations (PE), also known as implementation science, is the qualitative data component of the monitoring and evaluation platform of the MVP. For implementation projects, such as MVP, PEs are designed to address issues that impede the achievement of program objectives and the implementation of activities. Understanding the process of implementation is particularly relevant to the MVP. While the individual components of the package are of proven value, the systems necessary to support their integrated delivery in a diversity of settings are poorly understood. To address this, a portfolio of implementation science (or process evaluation) is conducted alongside the quantitative household impact surveys. These evaluations also help to distinguish between interventions that are inherently faulty (failure of intervention concept or theory) and those that were simply badly

delivered (implementation failure). Such assessments are increasingly recommended for evaluations of complex interventions (Guba and Lincoln, 1989; Oakley et al., 2004; Shiell et al., 2008).

The qualitative data from the PE serves 3 main objectives:

1. To document and describe the delivery systems developed over the 10 year project period, including sequencing of the core interventions undertaken in in each sector;
2. To outline major barriers and facilitators to implementation of the core interventions aimed to achieve MDGs;
3. To describe the learning vis-a-vis the implementation of MVP interventions over the 10 year project period.

4.11.1 Methodology

We will use two approaches in addressing the aforementioned PE objectives: *Key informant interviews* and *Community Focus Group Discussions (FGDs)*, which will be carried out with three levels of stakeholders. Sector-specific questionnaires for focus groups and individual interviews have been developed for the following informants:

1. MVP project staff and field implementers: Agriculture, Health, Education, Infrastructure, Community and Business Development Coordinators.
2. Village communities: Adult men and women, opinion leaders, teachers, local suppliers, village-based committee members.
3. Government partners: District and sub-district-level government officials, seconded government employees to MVP, field facilitators, and planning officers.

An estimated 20-25 key informant interviews will be conducted per MV study site with the aim to: 1) understand the major implementation barriers and facilitators of the interventions, 2) describe the MV delivery systems and their functionalities, 3) explore issues related to intervention synergies, externalities, and 4) explore issues related to transitioning to government. In addition to key informant interviews, a series of community focus group discussions (FGDs) will be conducted with the aim to: 1) understand the communitys experience of the MVP model of development over the 10 year project period, 2) obtain their opinions on the projects successes and failures, and 3) to assess their ability to adopt and maintain MVPs delivery systems and interventions after 2015.

The implementation of a complex project such as the MVP lends itself to questions related to its externalities via policy impacts and spread of services to nearby areas, the effects of its multi-sectoral synergies, and the effect of extra government attention. With regard to this, we will attempt to document these effects qualitatively through detailed interviews with communities within the MV project area and the district-level government. We will attempt to assess the local governance and the community organizational structure at the MV sites.

4.11.2 Recruitment and Sampling

We will sample using a criteria that aims to capture a diverse group of communities in terms of demographics, tribal and linguistic diversity, and those who live in the outer reaches of the project zone, as well as, those within the MV1 research zone. Due to limited resources, we will not conduct the PE in the comparison district areas and focus only on the clusters population to obtain participants. Ages of participants range from 16-59 years old. An informed consent process is administered verbally.

4.11.3 Data management and analysis

The content of qualitative interviews will be translated as necessary and transcribed. Analysis of qualitative data will involve thematic content analysis as well as critical appraisal where appropriate. A system of coding and memoing, will be facilitated by the use of qualitative analysis software (QSR International, 2008). Documentation templates will also serve to capture qualitative data and will allow for the inclusion of daily field notes from non-participatory observation in the community. From these data, the PE will generate a detailed description of the MVPs delivery systems, complemented by a mapping of the structural, financial and managerial components that were created in order to carry out the integrated package interventions; a description of the process of implementation alongside barriers and facilitators; a description of MVPs collaboration, in terms of planning, implementing and the process of handing over to governments and communities, and; finally a series of lessons learned that are inclusive of MVPs success and failures in its 10 year implementation.

4.11.4 Interpretation with Quantitative Data

Coupled with the survey results, the qualitative data better explains how interventions were designed and carried out, highlight implementation challenges and successes, and collate lessons learned in order to assist in replication and scale-up of the MVP. These combined quantitative and costing data will provide insights and lessons for replicability, scale-up and transfer of the MVP model to other contexts.

4.12 Transparency

We will post (via The Lancet) our small area estimation models, our propensity score model, and any additional information we have about the selection of MVs. We will post our ranking of candidate comparison areas, and ask the economic development and im-

pact evaluation communities to scrutinize our models and to comment on the candidate comparison villages. Though we may not be able to take all suggestions into account (as they may conflict), we will synthesize the feedback and come to a decision through an interactive process of external critique (Clemens and Demombynes, 2013).

Before end-line outcome data are available, but after the selection of candidate comparison villages, in the spirit of Humphreys et al. (2013) and Gelman and Carlin (2013), we will prepare and publicly release a design analysis using simulated fake data. This will help to set expectations regarding the final report. This design analysis will avoid some of the simplifying assumptions we make in our power calculations in Section 4.9.6.

We will make the analysis code and data public, so it can be reproduced and inspected by the scientific community.

4.13 Evaluation Timeline

The final 2015 MVP evaluation will consist of 4 major components with staggered releases of its findings. In July 2016, an MDG evaluation which will consist of an adequacy assessment of the MVP: "Did the MVP achieve the MDGs? will be made public. The findings from the remaining three evaluation components: 1) Impact Evaluation 2) Cost Assessment and 3) Process Evaluation will be released within a year, following the July 2016 adequacy assessment. These findings will be disseminated in high impact, peer-reviewed publications, project reports, implementation reviews, and presentations.

In addition to the MDG evaluation, the MVP will be releasing in 2016, a package of outputs of Lessons Learned from the project, including: books, articles, policy briefs, MV tools and the MV Field Guide. These outputs will also contain description of MVP's policy impacts, sustainability, replicability, and scalability of the MVP concept. Data from the cost assessment and process evaluation will be used to inform these subsequent outputs.

The survey data sets, costing datasets, and process evaluation summaries used in the final evaluation will be released through an online data archive. The datasets and analysis codes for all of the MVP data will be made available by January 1, 2017. A series of data analysis and dissemination workshops will be organized in the countries where the MVs are located to facilitate data sharing and utilization.

4.14 Ethical Issues

A number of important ethical issues have been addressed for the purposes of the study protocol:

1. **IRB approval:** All survey modules, questions and procedures employed as part of this assessment have undergone prior review and approval at Columbia Universitys Institutional Review Board (Protocol number AAAA8202) as well as approval by all host country IRBs.

2. **Community-level Informed consent:** Village leadership will be consulted prior to conducting assessments in all communities.

3. **Individual-level informed consent:** Informed consent will be obtained from all participating subjects. In the event of illiteracy witnessed verbal consent will be obtained prior to questionnaire administration. However, for all biological specimen collection (anemia and malaria) a signature or other form of written consent or mark will be obtained

4. **Minors:** As per the MVP protocol, adults will consent on behalf of survey respondents under 18 years old. Adults will give signed consent for blood specimens taken among under 5s.

5. **Non-coerced:** Explicit mention is made on the informed consents that assessments are not linked to any particular intervention being made available to individuals or

households at the time of survey or in the future.

6. **Confidentiality:** Confidentiality will be ensured through a number of mechanisms as per the existing quality assurance and data storage plans including: all source documents will be kept in locked cabinets at the villages; all data sent from the villages will be encrypted when transferred or stored; data will only be stored on limited access password protected computers; all database managers and investigators will have undergone IRB-approved training; all data will be anonymized prior to dissemination.

7. **Referral of the seriously ill:** All under 5s with fever, who are malnourished (by MUAC or child health cards) or who have moderate to severe anemia (Hemoglobin$< 110g/l$) will be immediately referred to the nearest health center for assessment.

8. **Stopping rules:** As this is not an assessment of unproven interventions, and as most interventions are delivered at the village rather than individual level, there are no stopping rules for the study.

4.15 Study Protocol Limitations and Future Areas of Research

There are a number of limitations that are important to underscore. These include:

1. **Study design:** As discussed in Sections 4.9.2 and A.7.3, the impact evaluation is severely weakened by the nonrandom design and lack of comparison area data at baseline. We believe we have outlined an approach that makes the best use of available data, adjusting for observable differences between treatment and control groups. The impact evaluation will be unavoidably subject to errors that will not be entirely quantifiable, but with all assumptions made clear, we hope that they can

be discussed transparently. Power is also a concern, because treatment is assigned at the cluster-level and no panel data in comparison areas prohibits adjustment for individual-level baselines.

2. **Both the intervention recipients and evaluation team are un-blinded to the intervention.** This has the potential to introduce interviewer or reporting bias and has been cited as a common challenge to community intervention trials (Donner and Klar, 2004; Sorenson et al., 1998). The use of standardized training of study personnel with clear standard operating procedures for field and data management systems is intended to minimize errors in survey enumeration, data capture, cleaning and analysis. During the informed consent procedure, it is made clear to respondents that participation in the evaluation has no bearing on the delivery of interventions at the household level.

3. **Lost to follow-up.** It is likely that those who leave a community may differ from those that do not. Random household replacement from baseline has been used to minimize the effect on statistical power. ITAD (2013) notes that high income earners may move to urban areas, which might cause us to underestimate program impact.

4. **There are no systems in place to monitor a number of important MDG outcomes.** These include HIV infection levels, TB incidence or malaria death rates. In addition, given the evaluation design, sample sizes are insufficient to detect cluster-level (i.e. MV-level) changes in other important indicators such as maternal mortality or adolescent fertility.

5. **Potential for Recall Bias.** Some indicators, such as child mortality rates, are themselves susceptible to recall bias. The longer back in history one measures, the greater the potential for error. In addition, non-surviving births are thought to be more frequently omitted than surviving births (Bicego and Ahmad, 1996). While this would cause mortality decline to be masked or underestimated, provided these errors are randomly distributed between intervention and comparison villages, we feel that

the overall effect of these errors on final risk ratios will be limited.

6. **Interviewee fatigue.** Those surveyed may get tired of answering survey questions, causing a deterioration in data quality.

7. **Definitive statements regarding mechanism of action will be difficult to make.** The process evaluation described in Section 4.11 will try to qualitatively reveal some of the mechanisms of action by studying differences between faulty intervention concepts and poor delivery (implementation failure). We believe that measuring these mechanisms (i.e. mediation analysis) is likely to be very difficult with this design. The regression framework proposed by Baron and Kenny (1986) relies on many strong assumptions (Green et al., 2010). Instead, we believe that measuring these causal mechanisms requires a separate causal study for each hypothesized mediator along the causal pathway from the MVP to outcomes. For example, if we are interested in the effect of the number of clinics built on health outcomes, we would want to design an experiment that randomly assigns different numbers of clinics to comparable areas, or to look for a natural experiment where the number of clinics vary. Examining separate causal studies is a large undertaking outside the scope of the final evaluation, so we leave it to future areas of research. Additionally, two very interesting questions remain largely unanswerable by this evaluation:

    - **We are unable to estimate synergistic effects.** As mentioned in Section 4.9.8, in the absence of an experimental design including arms with all possible treatment combinations, it is difficult to establish synergistic treatment interactions. It is also difficult to establish which component interventions are most effective, though it may be possible to find variables that are likely to be affected by one intervention and not the other, teasing apart which interventions work best (Duflo et al., 2008). This work will not be included in the this final evaluation, but may be in subsequent analyses. Moreover, it will be difficult and insufficient to survey the community to determine whether their perceived benefits were due to synergies that have taken place over the project's 10 year period.

While anecdotal evidence may be presented, this is insufficient to reliably assess synergistic effects. Estimating synergistic effects would answer one of the most important questions underlying the MVP (Sachs et al., 2004; Sachs, 2007): is the whole integrated package better than the sum of its parts? See Blattman (2007, 2009, 2010). It is an ambitious question, with sample size a major concern. It will be easier to meet the sample size requirements with individual-level randomization, for programs similar to the Ultra Poor Graduation program (Innovations for Poverty Action).

- **We are unable to attribute success to interventions versus institutions.** In addition to the interventions, institutions also play a key role in a population's ability to thrive. Attributing MV success to interventions versus institutions will be difficult. The process evaluation described in Section 4.11 will attempt to assess the extra government attention towards the project areas, and the governance at the sites. The cost assessment in Section 4.10 will highlight how government financial investments have trended over the project's duration - whether or not more spending has been covered by local government as the project winds down.

8. **External validity.** Extrapolating program effects beyond the villages operating thus far is problematic for a number of reasons. First, it is difficult to get estimates of the impact of the MVP at increasing scales. It seems likely that increasing the size of the population receiving health care would have a more positive effect on each individual in the society. Conversely, if a primary reason for success of the MVs is related to institutional accountability, scaling up may prove difficult, because the attention that enforces accountability is limited. In addition, extrapolating to time periods beyond 2005-2015 is difficult both because of global changes, and because lessons were learned in 2005-2015 that would likely be used in the next implementation of the MVP model.

9. **Externalities are difficult to assess.** In Section 4.9.7 we discuss the difficulty of measuring some types of externalities. We hope that the process evaluation described in Section 4.11 will be able to trace some of the externalities from the MVP.

10. **Sustainability is difficult to assess.** Without measurements in years following the MVP intervention, it will be difficult to assess the sustainability of its impacts. ITAD (2013) proposes assessment of sustainability via measurements 5 years after the end of the intervention. They also mention examining variables that have lasting impacts, such as child stunting. A definitive claim about sustainability of the impact in the years following implementation cannot be made with the data available in 2015-2016. The costing and process evaluations in Sections 4.10 and 4.11, respectively, will assess the ability of the sites to maintain MVP's delivery systems and interventions after 2015.

11. **Comparison to other projects with the same budget will be difficult to make.** An interesting, but largely unanswerable question given available data, is whether the MVP model is the best use of development money. For example, comparing the MVP to giving everyone in a village a cash transfer of equivalent value to running the whole project, including interventions and management (Haushofer and Shapiro, 2013; Blattman et al., 2013) To try to answer this question, one must find comparable areas (ideally via random assignment) and assign the MVP intervention to some, and the others receive a cash transfer. This would require a budget roughly double that of the MVP, to look at, for example, 20 villages, with ten randomly assigned to MVP and ten to cash transfers of equal value.

A. Appendices

A.1 A comparison of marginal and conditional models for capture-recapture data with application to human rights violations data

A.1.1 Model Fitting

To fit models considered in this paper we use Joseph Lang's `R` program `mph.fit`. (For documentation see: `http://www.divms.uiowa.edu/~jblang/mph.fitting/mph.fit.documentation.htm`) The program computes maximum likelihood estimates and model assessment statistics for the broad class of multinomial-Poisson homogeneous (MPH) and homogeneous linear predictor (HLP) models for contingency tables, (Lang, 2004, 2005).

For all models we assume a full multinomial sampling plan. We specify the temporary addition of a non-negative constant to the original data cell counts ($\epsilon = 0.1$) to avoid non-convergence problems caused by zero counts. At iteration five, after the algorithm has had time to move toward a non-boundary solution, the original counts are again used.

For the base simulations, the temporary addition of $\epsilon = 0.1$ is sufficient to alleviate the issue of sparsity (zero cell counts). However, in fitting the models to real Casanare data and the Casanare-inspired simulations, the issue of sparsity is severe and leads to more fitting difficulty. When maximum likelihood fitted values are non-existent due to zero counts, the log-scale moves toward negative infinity. Thus, the distance between the log-fitted values from iteration to iteration does not converge to zero, instead leveling off to some constant positive value. In this case, we instead use the distance between the score vector and zero to assess convergence. We also add a larger positive value ($\epsilon = 0.5$) to each cell count to avoid the boundary conditions in the beginning of the fitting algorithm.

For most of our simulation conditions, sparsity caused very few failures to fit models. However, for varied completeness, zero counts are very likely for all recording patterns where a list with low completeness has a capture, but the master list with high complete-

ness does not. This sparsity results in failures to fit models 1M and 1C in roughly 25% of simulations performed for varied completeness.

For fitting HLP models, the program `mph.fit` takes as input observed cell counts $\mathbf{n}$, whose expectations are $\boldsymbol{\mu}$, and fits a model of the form $L(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$. The models considered in this paper all specify links through the table probabilities $\boldsymbol{\pi}$. In most reasonable cases, GLLMs are HLP models (Lang, 2005). All models considered in this paper are HLP models.

Conditional model fitting

For the conditional models considered in this paper, we can separate the unobserved cell probability $\pi_{\mathbf{o}}$ from the observed cells and specify the model with only the $2^J - 1$ observed cell probabilities by deleting the first row of the design matrix $\mathbf{X}$, obtaining $L(\boldsymbol{\pi}_{obs}) = \log(\boldsymbol{\pi}_{obs}) = \mathbf{X}[-1,\ ]\boldsymbol{\lambda}$. To compute the MLE for N, we first compute the MLE for the missing cell probability as $\exp(\widehat{\lambda}_0) = \widehat{\pi}_{\mathbf{o}}$, and then $\widehat{N}_C = n/(1 - \widehat{pi}_{\mathbf{o}})$, the conditional model's estimate for the total population size.

Marginal model fitting

For the marginal model, the $\mathbf{A}$ and $\mathbf{C}$ of the GLLM formulation described in section* 3.3 are not identity matrices, as they are in the conditional model. The link is of the form $L(\boldsymbol{\pi}) = \mathbf{C}\log\mathbf{A}\boldsymbol{\pi} = \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{A}$ selects margins from $\boldsymbol{\pi}$ and $\mathbf{C}$ sets up contrasts to create marginal log odds and log odds ratios. Unlike in the conditional model, the unobserved cell probability does not separate into a distinct component of the design matrix $\mathbf{X}$. Thus, we must specify the model with the complete table of 2^J cells. We use the expectation maximization (EM) algorithm, where $n_{\mathbf{o}}$ is missing data (**?**).

To ensure convergence of the EM algorithm, we require it to run at least 50 iterations, and stop only when the deviance from the model fit stays stable (changing by at most 0.0005)

Table A.1: Casanare data results: interaction parameter estimates *[model fit]*.

Killings, $n = 2629$				
	QS models		QS2/QS3 models	
	Marginal	Conditional	Marginal	Conditional
3 NGOs, 4 govt	$\omega = 1.0$ *[2M]*	$\lambda = 0.7$ *[2C]*	$\begin{pmatrix}\omega_{NGO}\\ \omega_{govt}\\ \omega_{mix}\end{pmatrix} = \begin{pmatrix}3.9\\ 1.7\\ 0.6\end{pmatrix}$ *[3M]*	$\begin{pmatrix}\lambda_{NGO}\\ \lambda_{govt}\\ \lambda_{mix}\end{pmatrix} = \begin{pmatrix}2.7\\ 1.1\\ 0.1\end{pmatrix}$ *[3C]*
Collapsed NGOs, 4 govt	$\omega = 1.5$ *[4M]*	$\lambda = 0.7$ *[4C]*	$\begin{pmatrix}\omega_{govt}\\ \omega_{mix}\end{pmatrix} = \begin{pmatrix}1.8\\ 0.6\end{pmatrix}$ *[5M]*	$\begin{pmatrix}\lambda_{govt}\\ \lambda_{mix}\end{pmatrix} = \begin{pmatrix}1.1\\ 0.2\end{pmatrix}$ *[5C]*

Disappearances, $n = 867$				
	QS models		QS2/QS3 models	
	Marginal	Conditional	Marginal	Conditional
2 NGOs, 5 govt	$\omega = 0.7$ *[2M]*	$\lambda = 0.6$ *[2C]*	$\begin{pmatrix}\omega_{NGO}\\ \omega_{govt}\\ \omega_{mix}\end{pmatrix} = \begin{pmatrix}2.2\\ 0.7\\ 0.4\end{pmatrix}$ *[3M]*	$\begin{pmatrix}\lambda_{NGO}\\ \lambda_{govt}\\ \lambda_{mix}\end{pmatrix} = \begin{pmatrix}2.2\\ 0.7\\ 0.5\end{pmatrix}$ *[3C]*
Collapsed NGOs, 5 govt	$\omega = 0.7$ *[4M]*	$\lambda = 0.6$ *[4C]*	$\begin{pmatrix}\omega_{govt}\\ \omega_{mix}\end{pmatrix} = \begin{pmatrix}0.8\\ 0.6\end{pmatrix}$ *[5M]*	$\begin{pmatrix}\lambda_{govt}\\ \lambda_{mix}\end{pmatrix} = \begin{pmatrix}0.7\\ 0.5\end{pmatrix}$ *[5C]*

for at least three iterations. If this stabilization fails to be achieved within 600 iterations, we report a failure of the algorithm. As a starting value for the missing cell count, we use the predicted missing cell count from the conditional model.

A.1.2 Casanare Data Analysis

Here we include tables to display the interaction parameter estimates and goodness of fit statistics from models 2M, 2C, 3M, 3C, 4M, 4C, 5M, 5C fit to the Casanare data.

A.1.3 Simulation conditions

For low, medium and high completeness, we take $\text{logit}(p_1), ..., \text{logit}(p_J)$ from $\alpha - \sigma$ to $\alpha + \sigma$ evenly spaced, with $\sigma = 1$. For low completeness $\alpha = -1.5$, for medium $\alpha = 0$,

Table A.2: Casanare data results: deviance (G^2) and degrees of freedom (df) *[model fit]*.

Killings, $n = 2629$					
	QS models		QS2/QS3 models		two-way model
	Marginal	Conditional	Marginal	Conditional	Conditional
3 NGOs, 4 govt	$G^2 = 867, df = 118$ *[2M]*	$G^2 = 868, df = 118$ *[2C]*	$G^2 = 460, df = 116$ *[3M]*	$G^2 = 501, df = 116$ *[3C]*	$G^2 = 69, df = 98$ *[1C]*
Collapsed NGOs, 4 govt	$G^2 = 459, df = 24$ *[4M]*	$G^2 = 492, df = 24$ *[4C]*	$G^2 = 349, df = 23$ *[5M]*	$G^2 = 375, df = 23$ *[5C]*	$G^2 = 31, df = 15$ *[1C]*

Disappearances, $n = 867$					
	QS models		QS2/QS3 models		
	Marginal	Conditional	Marginal	Conditional	
2 NGOs, 5 govt	$G^2 = 1200, df = 118$ *[2M]*	$G^2 = 1180, df = 118$ *[2C]*	$G^2 = 1170, df = 116$ *[3M]*	$G^2 = 1146, df = 116$ *[3C]*	$G^2 = 121, df = 98$ *[1C]*
Collapsed NGOs, 5 govt	$G^2 = 1020, df = 55$ *[4M]*	$G^2 = 1004, df = 55$ *[4C]*	$G^2 = 1017, df = 54$ *[5M]*	$G^2 = 997, df = 54$ *[5C]*	$G^2 = 57, df = 41$ *[1C]*

and for high $\alpha = 1.5$. Note that the logit's symmetry around zero makes list completeness values in the medium case symmetric around 0.5. We see symmetry in the three-way odds ratios (ORs) because each three-way OR has a "mirror image" by taking the three lists with completeness values reflected over 0.5. Also, because the α for low and high completeness are opposite in sign, for $j = 1, ..., J$, we have that $\text{logit}(p_j^{low}) = -1.5 - 2\sigma/J = -\text{logit}(p_{J-j+1}^{high})$, so in fact the set of completeness values for low are $\{1 - p_j^{high}\}$. We see that three-way ORs for low completeness are the mirror image across zero of the three-way ORs for the high, while the two-way ORs are exactly identical for the low and high completenesses.

For $J = 4$, list recording probabilities, i.e. completeness values, are $\mathbf{p} = (0.08, 0.14, 0.24, 0.38)$ for low, $(0.27, 0.42, 0.58, 0.73)$ for medium, $(0.62, 0.76, 0.86, 0.92)$ for high, and $(0.08, 0.18, 0.38, 0.88)$ for varied. For $J = 6$, list recording probabilities are $(0.08, 0.11, 0.15, 0.21, 0.29, 0.38)$ for low, $(0.27, 0.35, 0.45, 0.55, 0.65, 0.73)$ for medium, $(0.62, 0.71, 0.79, 0.85, 0.89, 0.92)$ for high, and $(0.08, 0.12, 0.18, 0.27, 0.38, 0.88)$ for

varied. For our Casanare-inspired simulations we use list completeness values $\mathbf{p} = (0.03, 0.04, 0.06, 0.08, 0.10, 0.14, 0.18)$, using $\alpha = -2.5$. From $\mathbf{p}$ we compute $\boldsymbol{\theta} = (\theta_1, ..., \theta_J)$ using the multivariate normal distribution.

For base simulations, we generate data from the tetrachoric correlation model with exchangeable correlation d. We explore possible d values such that the correlation matrix remains positive-definite. This restricts $-1/(J-1)d < 1$. We avoid exploring values of d near the boundaries of this interval, as fitting difficulties increase when we approach non-positive-definite correlation matrices. Note that $d = 1$ represents J identical lists, and for nearly identical lists, capture-recapture has limited use. For $J = 4$, we explore d values ranging from -0.23 to 0.7 and for $J = 6$ we explore d values from -0.11 to 0.7. Note that moving from 4 to 6 lists, we cannot have lists be as negatively correlated.

A.1.4 More Simulation Results

In this section we include extra figures describing results from the simulations. To characterize the simulated data from the tetrachoric model, we plot information about cell means $\{\mu_{\mathbf{k}}\}$ in Figures A.1 and A.2. All models we fit assume no three-way and higher marginal or conditional interactions. To assess model misspecification, we plot log three-way odds ratios (ORs) as boxplots. We plot two-way ORs, with heterogeneity in these ORs reflecting model misspecification in the QS models 2M and 2C that assume these ORs to be homogeneous across pairs of lists. For the conditional model, no three-way and higher interactions implies that for any pair of lists, their OR is homogeneous across recording patterns in other lists. QS and no higher-order interactions is equivalent to heterogeneous two-way conditional ORs being equal. However, in the marginal model, equal two-way marginal ORs does not imply no higher-order marginal interactions.

We see that both marginal and conditional two-way ORs are negative for negative correlation on the tetrachoric scale (d) and positive for positive d values. For low and varied completeness, three-way ORs are mostly negative, for high completeness mostly positive,

and for medium they are symmetric about zero (see Section A.1.3). For d values further from zero, magnitudes of three-way ORs are higher and there is more variation in two-way ORs. Patterns become more pronounced moving from $J = 4$ to $J = 6$ lists. In the bottom row of Figure A.2 we plot the mean cell count for the missing cell ($\mu_{\mathbf{o}}$), which is larger when the lists are less complete or association between lists increases.

Figures A.3, A.4, A.5, and A.6 show $|\widehat{N}_M - N| - |\widehat{N}_C - N|$ and $\widehat{N}_M - \widehat{N}_C$ for estimates of N from the marginal models 1M, 2M versus the conditional models 1C, 2C for $J = 4$ and $J = 6$.

Corresponding to Figure 2.2 in the main paper, in Figures A.7, A.8, and A.9 we plot the distribution of $\widehat{N} - N$ as boxplots, and we plot the RMSE computed across the simulations. We use solid lines and filled black boxplots for QS conditional model 2C and the low sample coverage estimator, and dashed lines and empty gray boxplots for the QS marginal model 2M and high sample coverage estimator.

In Figures A.10 and A.11 we plot the coverage of the 95% profile likelihood intervals for our base simulations and Casanare-inspired simulations respectively. We discuss the results in Section 2.5.1 in the main paper.

Figures A.12 and A.13 show results from fitting the QS models 2M and 2C to the Casanare-inspired simulations, corresponding to Figures 2.3 and 2.4 in the main paper, which showed results from the QS3 models 3M and 3C.

A.1.5 Data Descriptives

Here we include some summaries of the data.

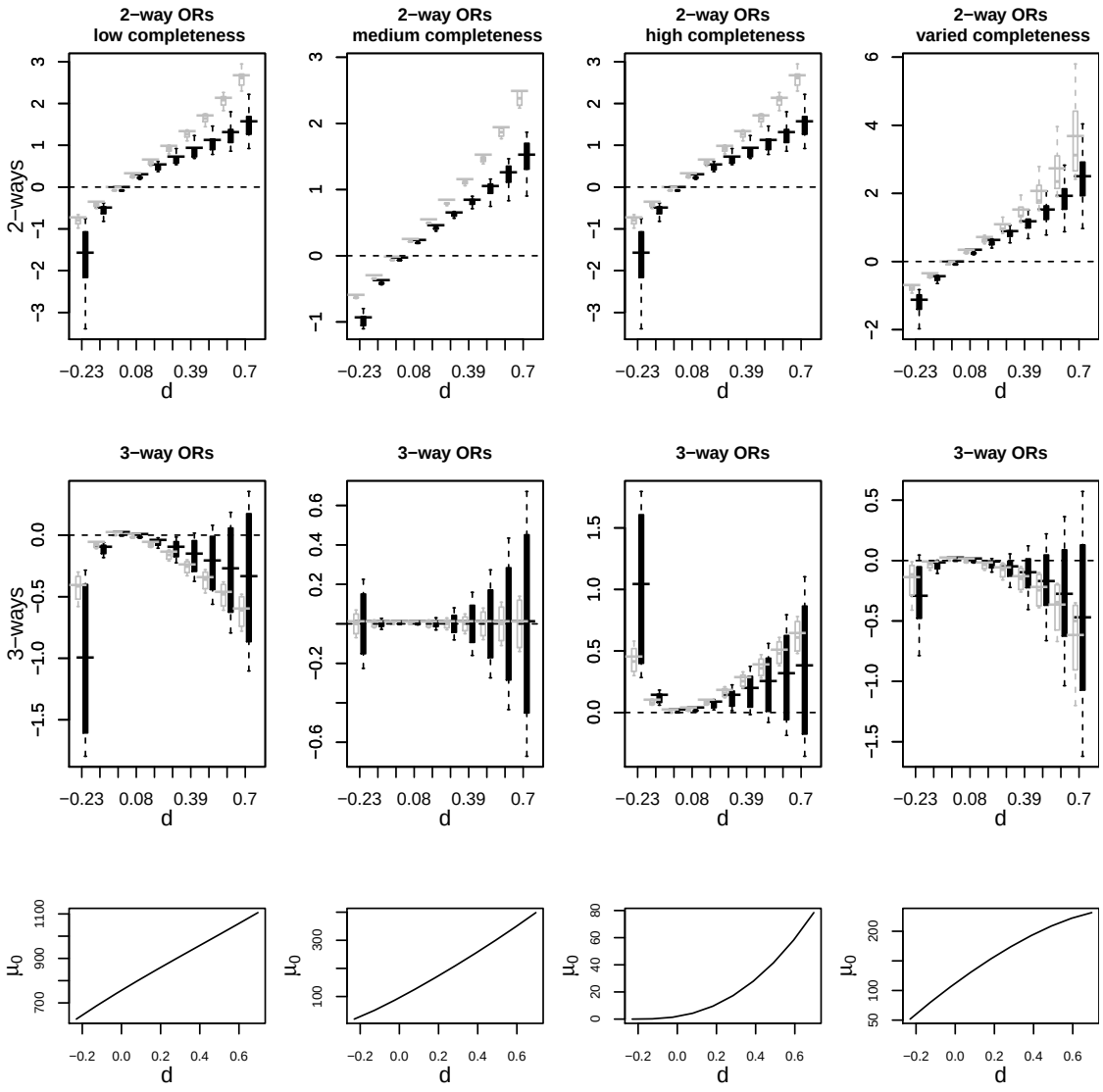


Figure A.1: Plots show information about the generated data for $J = 4$ lists. We use filled black boxplots for the conditional ORs, and empty gray boxplots for the marginal ORs.

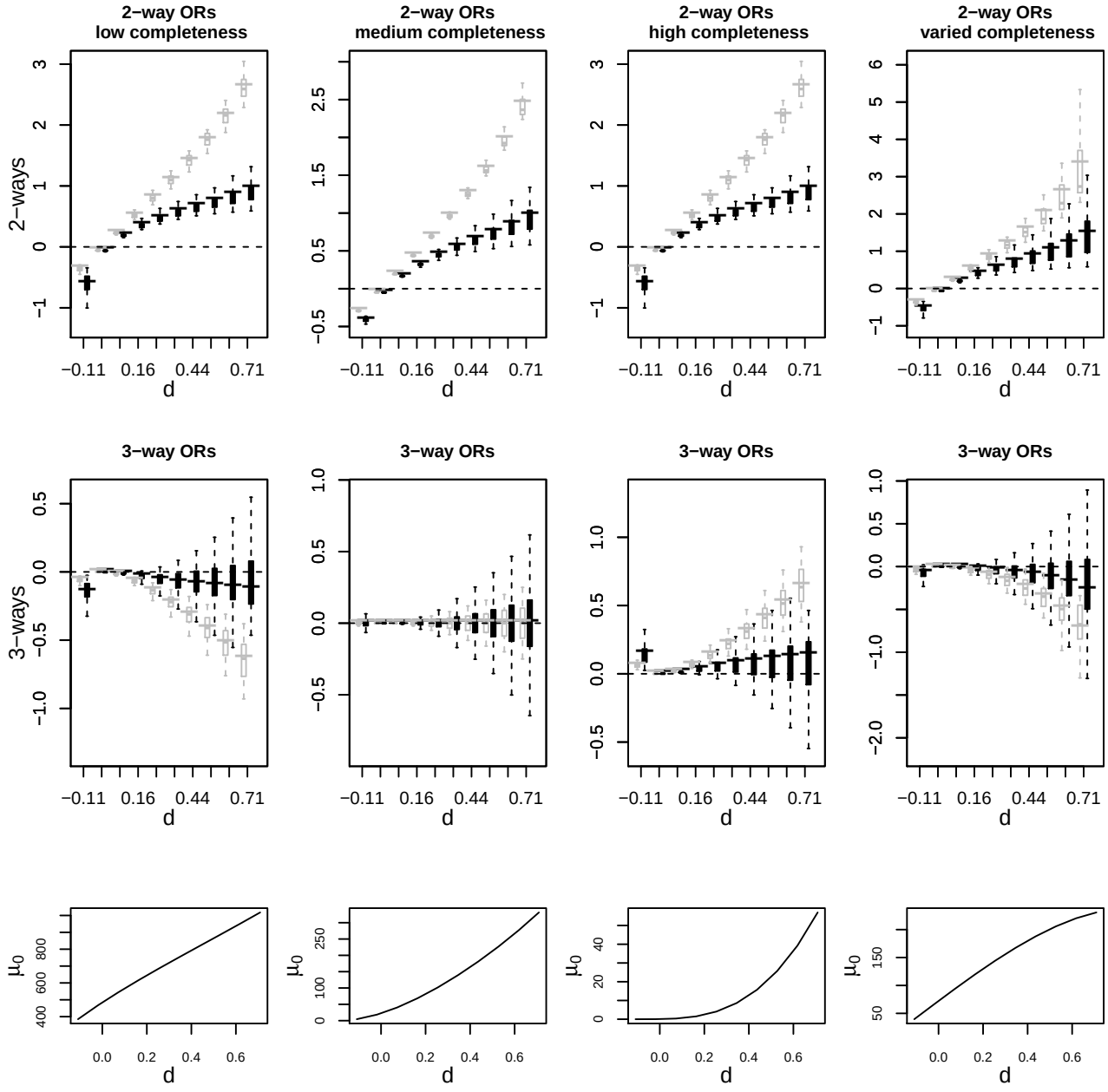


Figure A.2: Plots show information about the generated data for $J = 6$ lists across the exchangeable correlation on the tetrachoric scale (d). We use filled black boxplots for the conditional ORs, and empty gray boxplots for the marginal ORs. We plot the mean cell count for the missing cell ($\mu_{\mathbf{o}}$) in the bottom row.

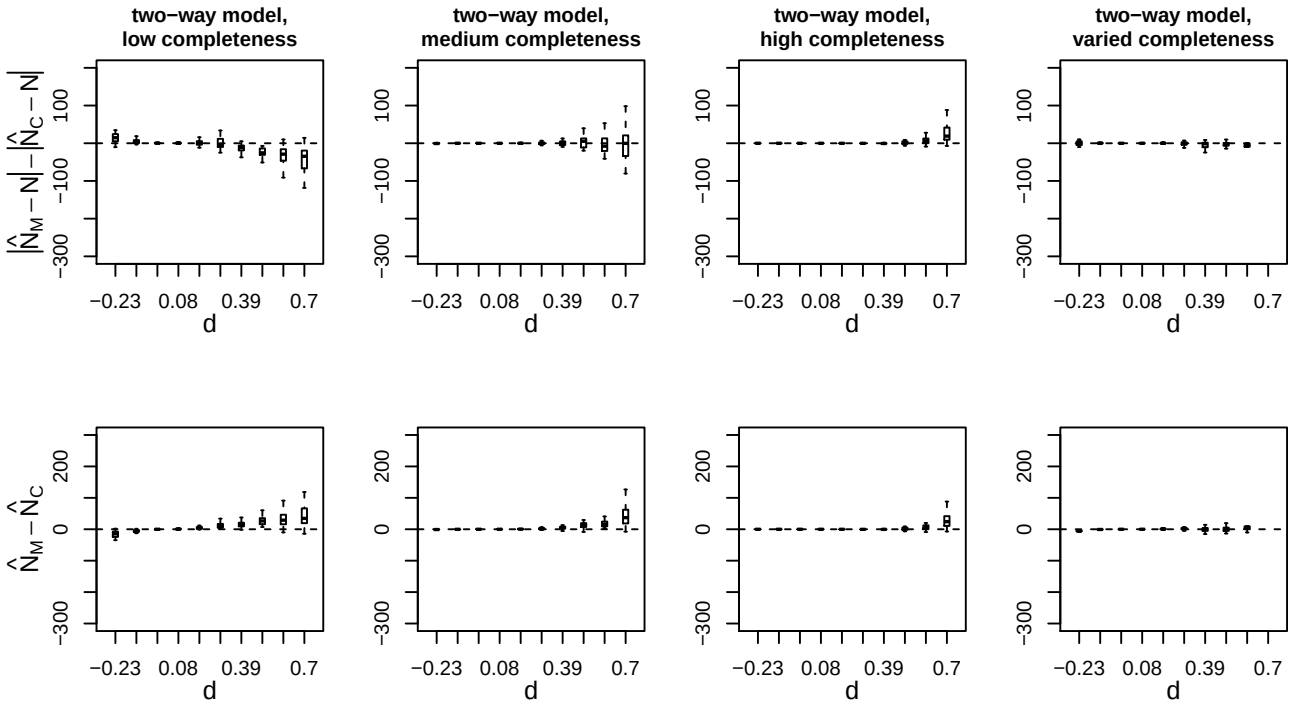


Figure A.3: Plots show $|\widehat{N}_M - N| - |\widehat{N}_C - N|$ and $\widehat{N}_M - \widehat{N}_C$ for estimates of N from the marginal heterogeneous two-way model 1M versus the conditional heterogeneous two-way model 1C fit to simulated data. Data simulated have $N = 2000$, $J = 4$ lists, and an exchangeable tetrachoric correlation structure with correlation d.

Table A.3: Distribution of the number of times a record appears on the lists, i.e. the number of captures.

Disappearances	
Number of captures	Frequency
1	500
2	202
3	79
4	66
5	18
6	2
7	2
8	1
9	1
10	1

Killings	
Number of captures	Frequency
1	1847
2	578
3	168
4	23
5	11
7	2

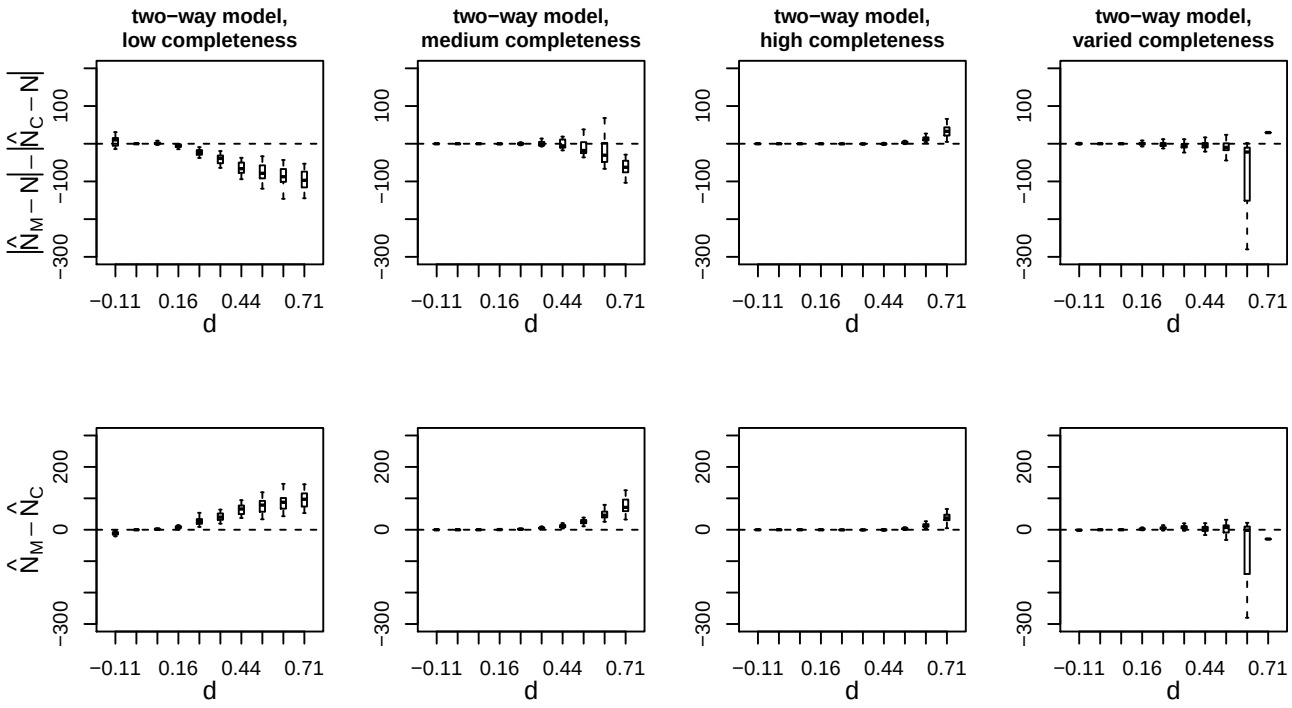


Figure A.4: Plots show $|\widehat{N}_M - N| - |\widehat{N}_C - N|$ and $\widehat{N}_M - \widehat{N}_C$ for estimates of N from the marginal heterogeneous two-way model 1M versus the conditional heterogeneous two-way model 1C fit to simulated data. Data simulated have $N = 2000$, $J = 6$ lists, and an exchangeable tetrachoric correlation structure with correlation d.

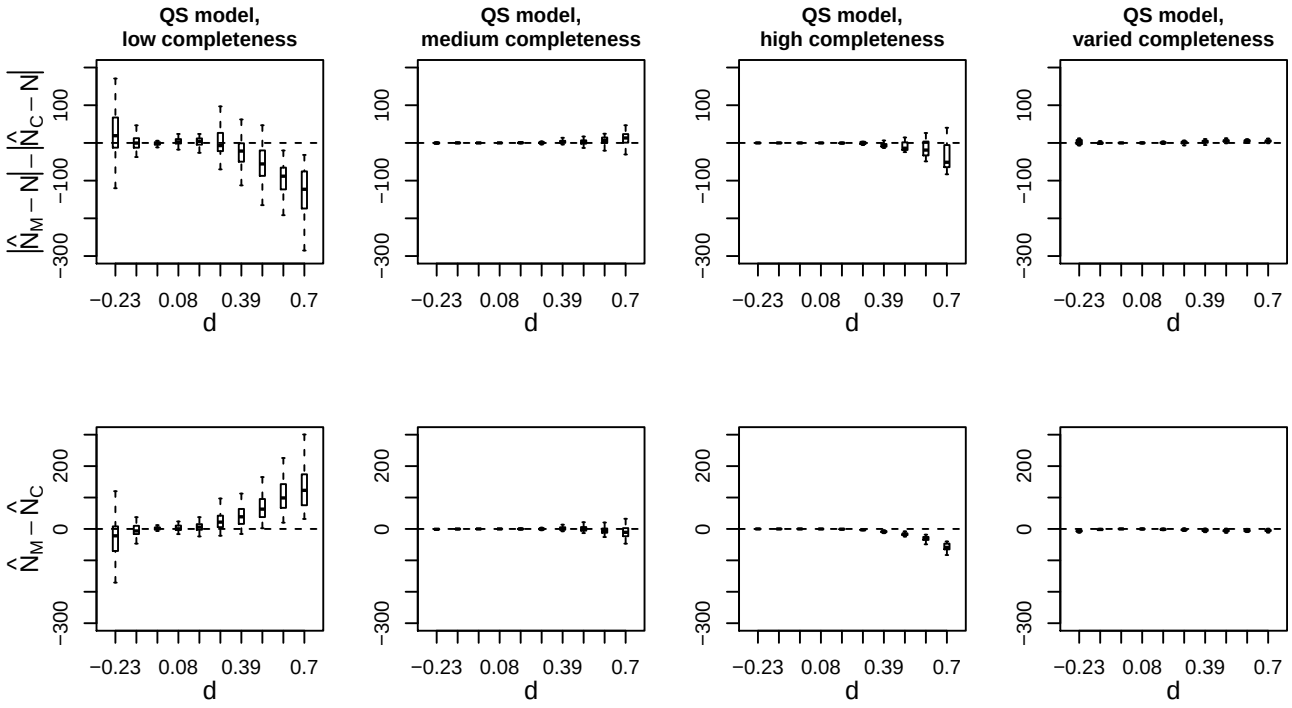


Figure A.5: Plots show $|\widehat{N}_M - N| - |\widehat{N}_C - N|$ and $\widehat{N}_M - \widehat{N}_C$ for estimates of N from the marginal QS model 2M versus the conditional QS model 2C fit to simulated data. Data simulated have $N = 2000$, $J = 4$ lists, and an exchangeable tetrachoric correlation structure with correlation d.

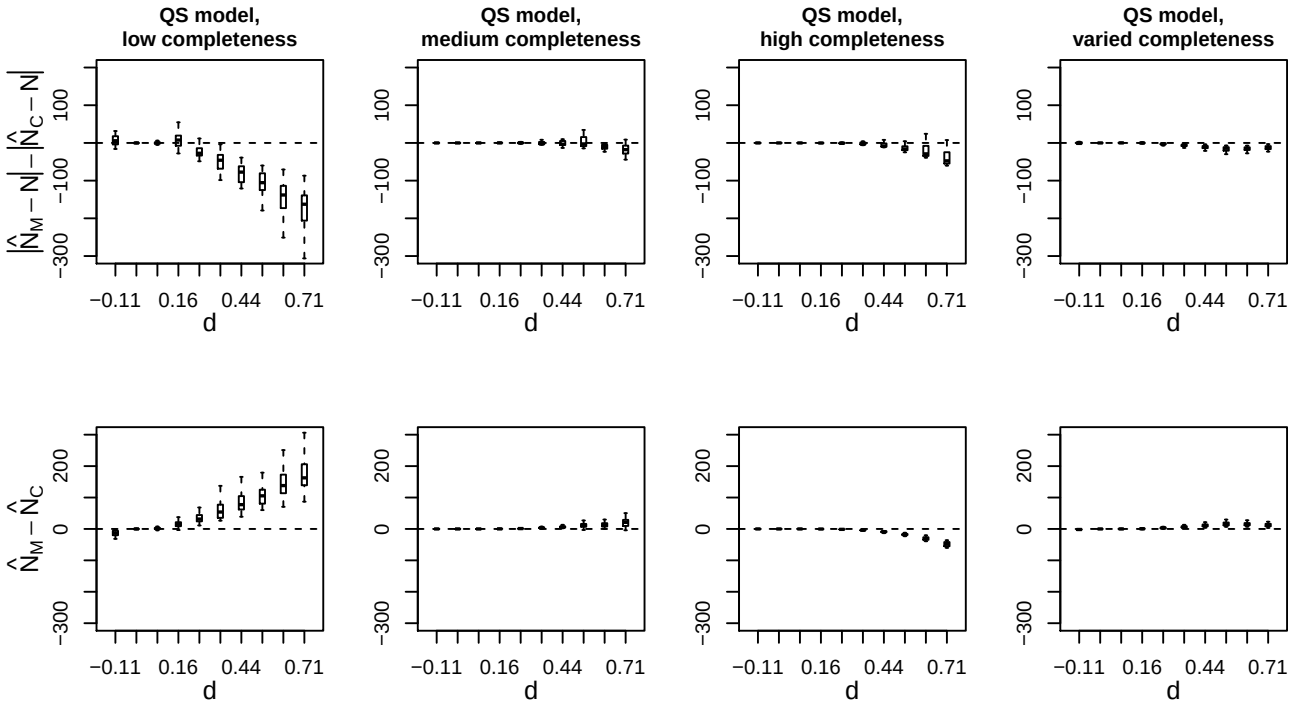


Figure A.6: Plots show $|\widehat{N}_M - N| - |\widehat{N}_C - N|$ and $\widehat{N}_M - \widehat{N}_C$ for estimates of N from the marginal QS model 2M versus the conditional QS model 2C fit to simulated data. Data simulated have $N = 2000$, $J = 6$ lists, and an exchangeable tetrachoric correlation structure with correlation d.

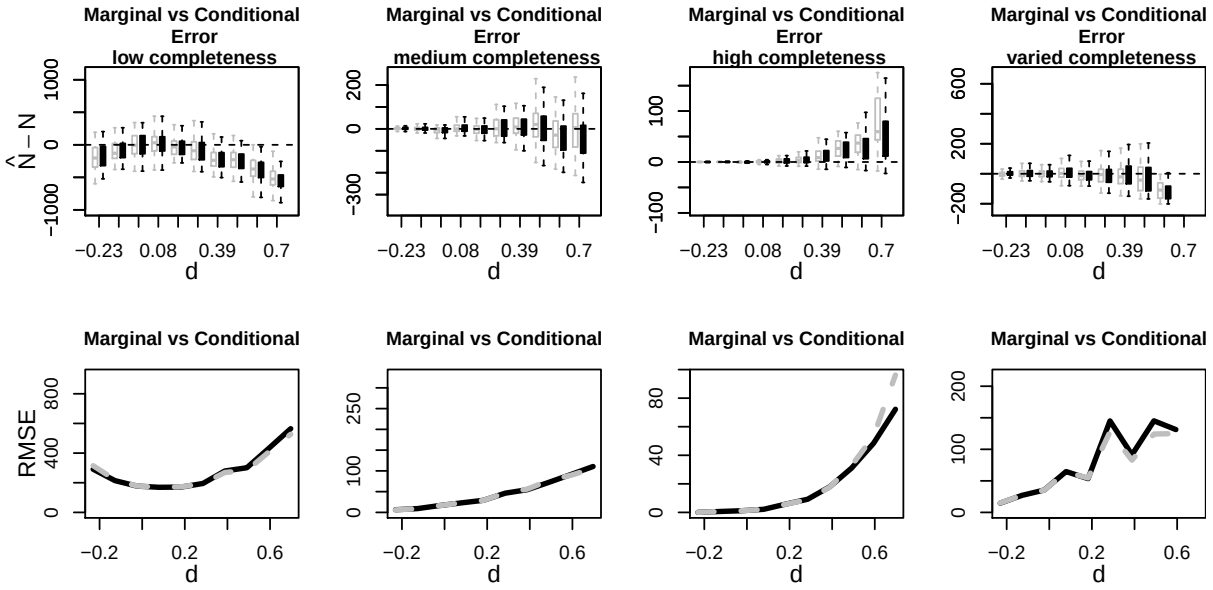


Figure A.7: Results from the base simulations: $N = 2000$ true total population size, $J = 4$ lists, exchangeable correlation structure. In row one we plot the distribution of $\widehat{N} - N$ across simulations (as boxplots), and in row two we plot the RMSE, where we note that almost all the MSE is attributable to the bias rather than variance. We use solid lines and filled black boxplots for heterogeneous two-way conditional model 1C and dashed lines and empty gray boxplots for the heterogeneous two-way marginal model 1M.

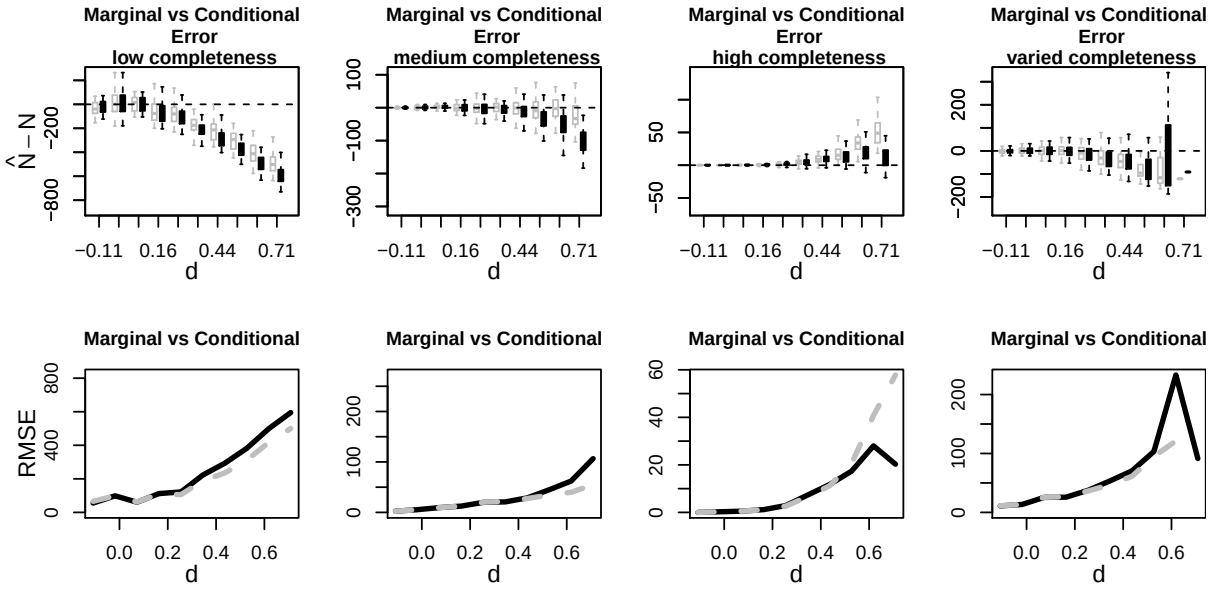


Figure A.8: Results from the base simulations: $N = 2000$ true total population size, $J = 6$ lists, exchangeable correlation structure. In row one we plot the distribution of $\widehat{N} - N$ across simulations (as boxplots), and in row two we plot the RMSE, where we note that almost all the MSE is attributable to the bias rather than variance. We use solid lines and filled black boxplots for heterogeneous two-way conditional model 1C and dashed lines and empty gray boxplots for the heterogeneous two-way marginal model 1M.

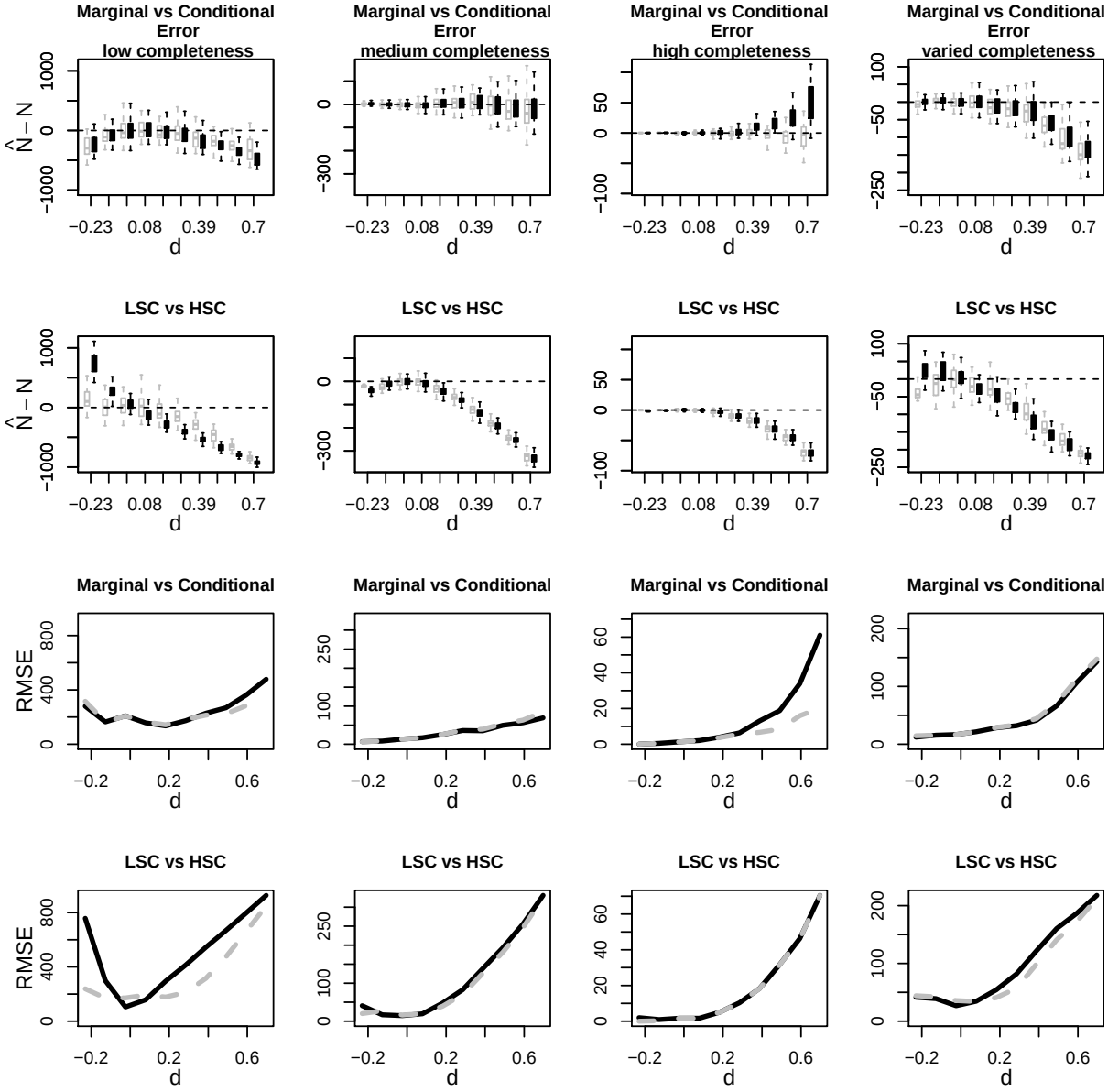


Figure A.9: Results from the base simulations: $N = 2000$ true total population size, $J = 4$ lists, exchangeable correlation structure. In rows one and two, we plot the distribution of $\widehat{N} - N$ across simulations (as boxplots), and in rows three and four we plot the RMSE, where we note that almost all the MSE is attributable to the bias rather than variance. We use solid lines and filled black boxplots for QS conditional model 2C and the low sample coverage estimator, and dashed lines and empty gray boxplots for the QS marginal model 2M and high sample coverage estimator.

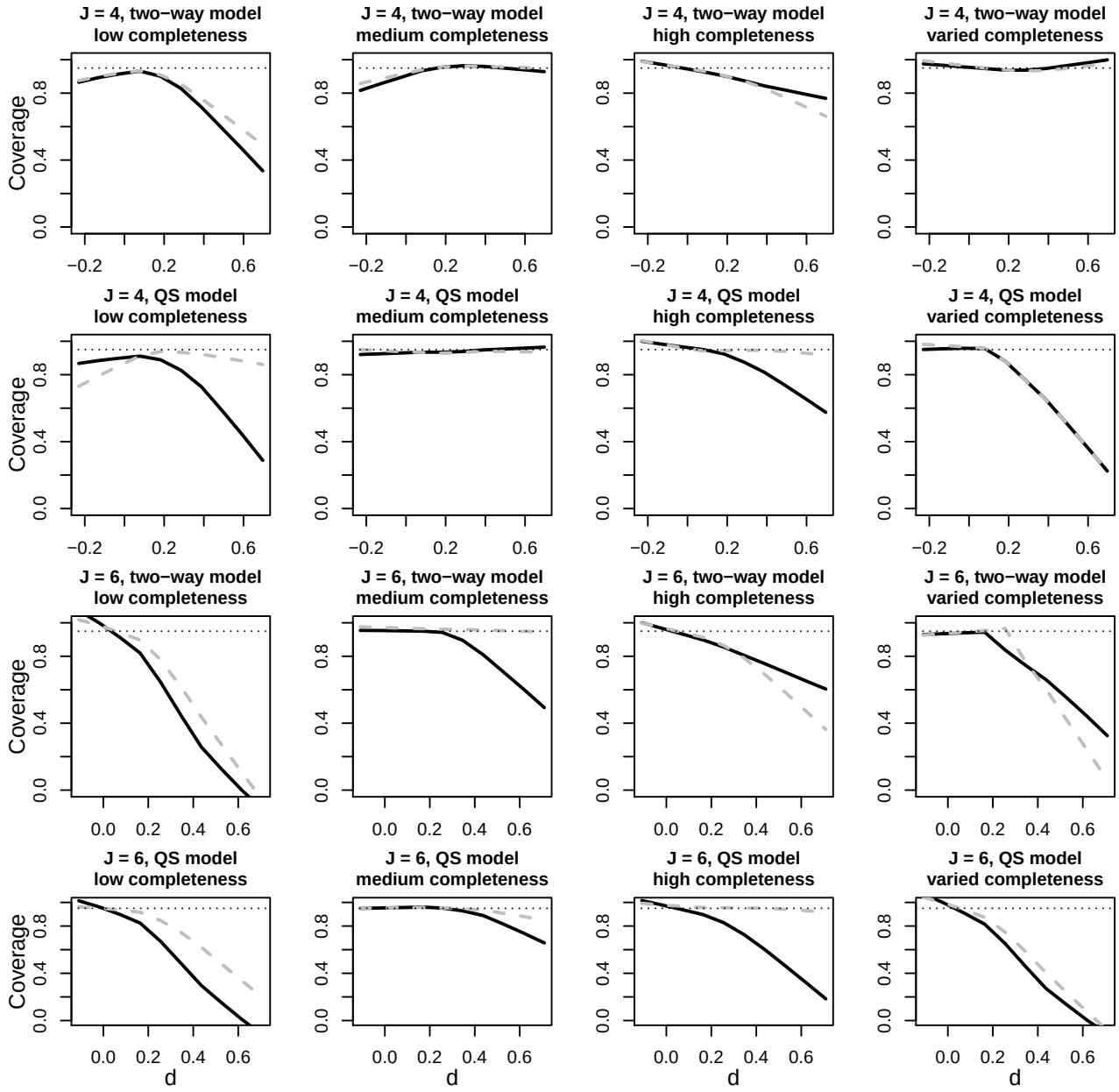


Figure A.10: Coverage of the 95% profile confidence intervals for the base simulations. We use solid lines for the conditional models 1C and 2C, and dashed lines for the marginal models 1M and 2M.

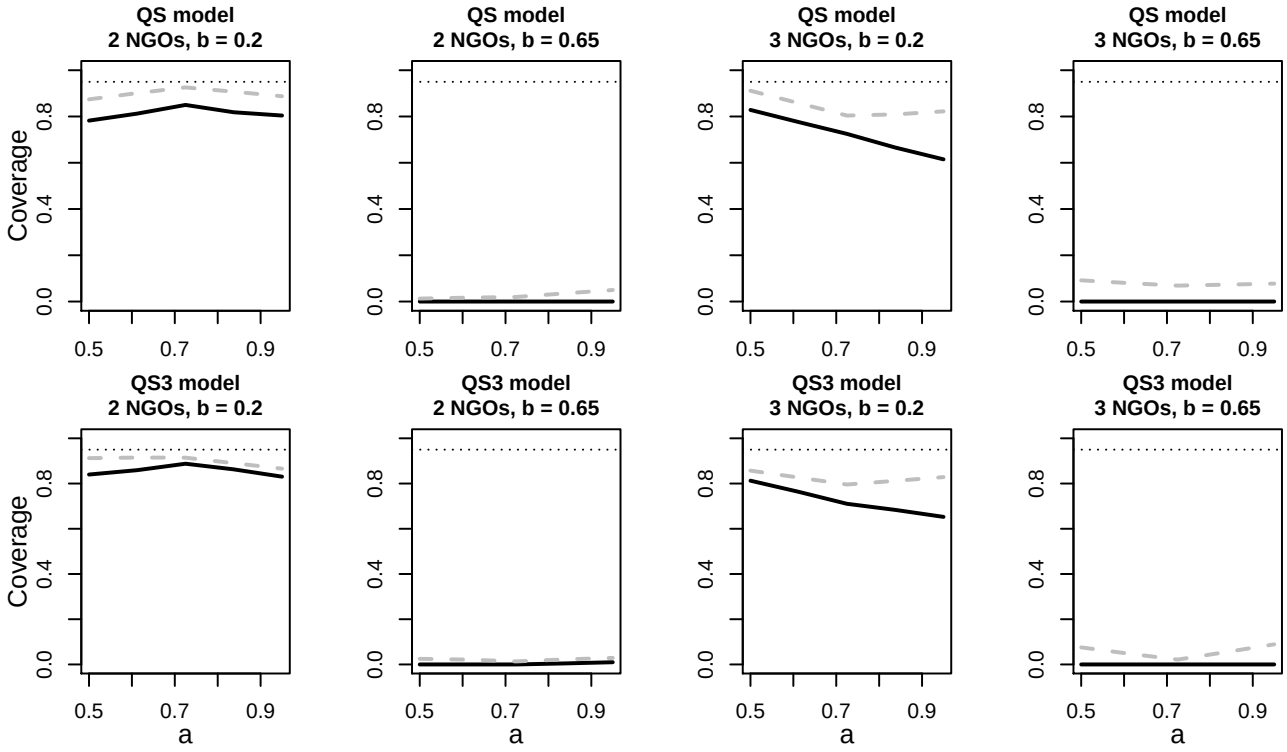


Figure A.11: Coverage of the 95% profile confidence intervals for the Casanare-inspired simulations. We use solid lines for the conditional models 2C and 3C, and dashed lines for the marginal models 2M and 3M.

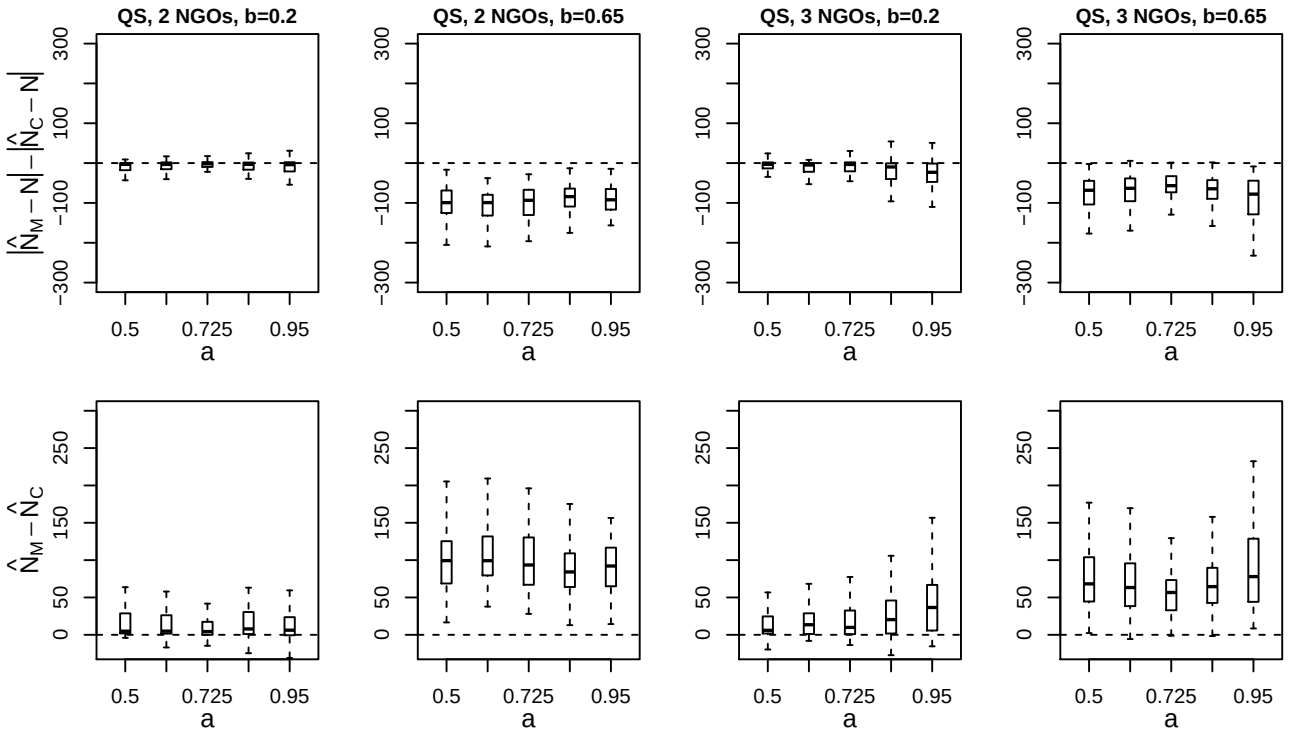


Figure A.12: The plot shows $\widehat{N}_M - \widehat{N}_C$ and $|\widehat{N}_M - N| - |\widehat{N}_C - N|$ for estimates of N from the marginal QS model 2M versus the conditional QS model 2C fit to simulated data. Data simulated have $N = 2000$, $J = 7$ lists, and a block tetrachoric correlation structure by list type, where b is government list association, a is NGO association, and c is association across type, which we set at $c = b/2$.

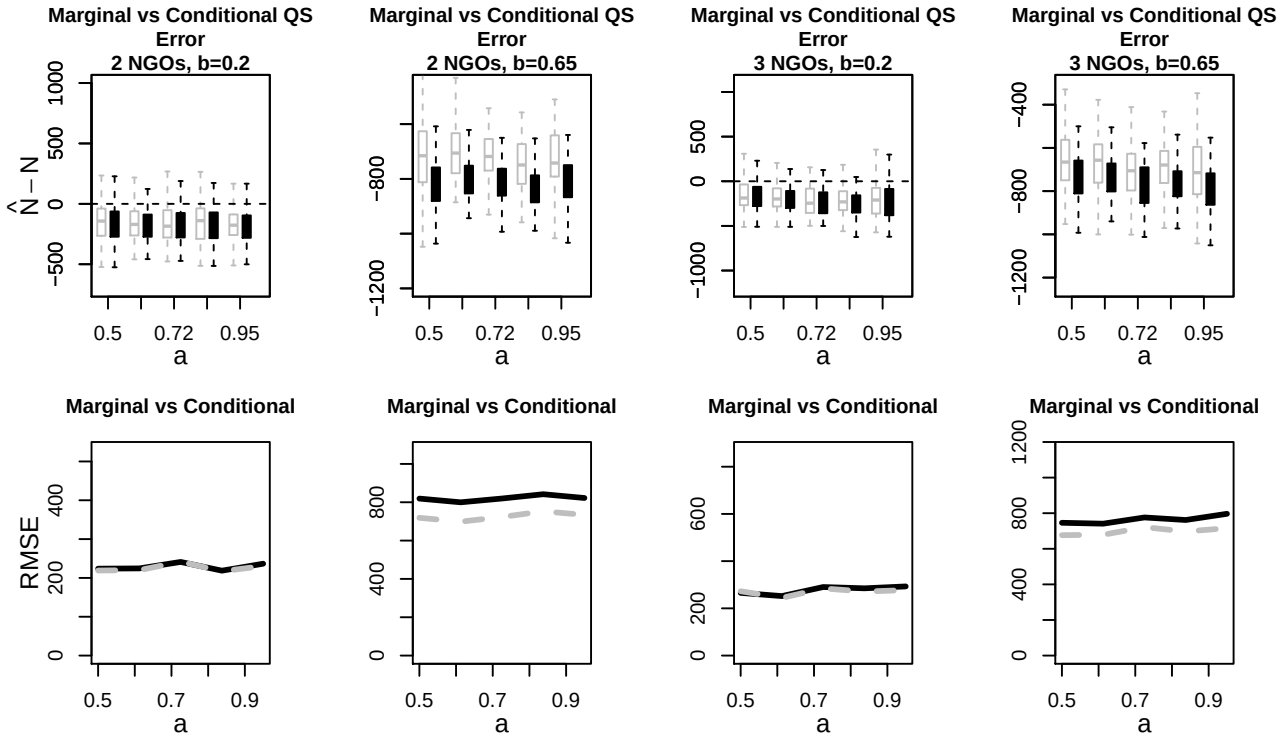


Figure A.13: In row one we plot the distribution of $\widehat{N} - N$ across simulations (as boxplots), and in row two we plot the RMSE, where we note that almost all the MSE is attributable to the bias rather than variance. We use solid lines and filled black boxplots for QS conditional model 2C and dashed lines and empty gray boxplots for the QS marginal model 2M. Data simulated have $N = 2000$, $J = 7$ lists, and a block tetrachoric correlation structure by list type, where b is government list association, a is NGO association, and c is association across type, which we set at $c = b/2$.

Table A.4: The total records column indicates how many records are in each list after de-duplication within lists. Next, we give totals in 1998-2007, totals in 1998-2007 where records without date or municipality information were dropped, then broken down by killings and disappearances. "Uniques" counts records found only on that list, not in any other. Finally, we include the type of list (government or non-government), notes, and abbreviation.

Organization name	Total	1998-2007	with yr, muni	Kill	Disp	Uniques	type	Notes	abbrev
National Institute of Forensic Medicine Deaths	2168	2085	1878	1874	4	1420	govt		IMLM
Prosecutor General list of the Disappeared	1313	659	623	24	599	387	govt		FDC
Human Rights Observatory of the Vice Presidency	528	501	501	501	0	284	govt		VP
National Police - DIJIN	825	825	825	825	0	221	govt		PN0
CINEP	338	274	267	138	129	97	NGO		CIN
Fondelibertad	332	312	304	32	272	67	govt	manages money for Gaula	FON
Colombian Commission of Jurists	250	217	214	160	54	51	NGO		CCJ
Colombia-Europe-US Coordination	77	72	72	72	0	30	NGO		CCE
National Institute of Forensic Medicine Disappearances	193	172	153	7	146	10	govt		IMLD
Prosecutor General of Santa Rosa	169	163	151	5	146	4	govt		FSR
Families of Victims Organizations	52	51	51	4	47	1	NGO		FAM
Gaula (Anti-extortion Unit)	128	111	110	19	91	1	govt	navy	GAU
Technical Investigative Body of the Prosecutor Generals Office	36	36	36	5	31	0	govt		CTI
Equitas	28	23	22	1	21	0	NGO		EQU
Pais Libre	9	9	9	1	8	0	NGO		PL

Matching rules

From Guberek et al. (2010), the following rules were used to match the data by the Human Rights Data Analysis Group (HRDAG). HRDAG grouped together multiple records on the same victim into a "match group." When the records were not exact matches on violation type, contradictions were resolved by the rules:

- If at least one record in the match group was a killing, the group was determined to be a killing.
- If at least one record was a disappearance and there was no killing record in the group, the group was considered a disappearance.
- Records of detentions, hostages and extortive kidnappings were only kept if they were matched with a record of killing or disappearance, the rest were dropped from the data.

The data were matched by two human matchers, who showed a high rate of agreement, with 280 pairs of records matched by the first matcher but not the second, 149 pairs matched by the second matcher but not the first, and 4,389 pairs of records matched by both of them (Guberek et al., 2010).

A.1.6 Acknowledgements

The authors thank HRDAG for the inspiration for this work, their cleaned and matched data, and subject-matter knowledge. In particular, we thank Megan Price and Patrick Ball for guidance in the use of capture-recapture methods for human rights data. The authors thank Professors Peter van der Heijden and Alan Agresti for input and expertise in capture-recapture methods. The authors are grateful to the Editors and referees for numerous suggestions that significantly improved the paper. This material is based

upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1144152.

A.2 Population Size Estimation with Inactive Lists: Hierarchical mixture models and Missing Data with Application to Armed Conflict Data

A.2.1 The MCMC Computation

Convergence was evaluated informally by looking at trace plots, and was obtained after around 10,000 samples.

Let $\Sigma_j(\boldsymbol{\gamma}_j, \rho, \tau^2)$ be the variance of the prior distribution $\boldsymbol{\lambda}_j|\boldsymbol{\gamma}_j, \mu_j, \rho, \tau^2$ in 3.5, then $\Sigma_j(\boldsymbol{\gamma}_j, \rho = 0, \tau^2) = \Sigma_j(\boldsymbol{\gamma}_j, \tau^2)$ is the variance of the prior distribution $\boldsymbol{\lambda}_j|\boldsymbol{\gamma}_j, \mu_j, \tau^2$ in 3.3. We abbreviate both as Σ_j. For both models, the likelihood of the complete data is

$$\mathcal{L}_{complete}(\boldsymbol{\lambda}, \boldsymbol{\omega}, \mathbf{N}, \boldsymbol{\gamma}) \equiv \prod_{t=1}^{T} \underbrace{\binom{N^{(t)}}{\{n_{\mathbf{k}}^{(t)}\}_{\mathbf{k}}} \prod_{\mathbf{k}} \boldsymbol{\pi}_{\mathbf{k}}^{(t)}(\boldsymbol{\lambda}_t, \boldsymbol{\omega}, \boldsymbol{\gamma}_t)^{n_{\mathbf{k}}^{(t)}}}_{\mathcal{L}_t}.$$

The Gibbs sampling steps - Hierarchical mixture model

Then the joint distribution

$$
\begin{aligned}
& p\left(\{N^{(t)}\}_t, \{\lambda_{j,t}\}_{j,t}, \boldsymbol{\omega}, \{\gamma_{j,t}\}_{j,t}, \{\mu_j\}_j, \tau^2, \{n_{\mathbf{k}}^{(t)}\}_{\mathbf{k}\neq\mathbf{o},t}\right) \\
&= p\left(\{N^{(t)}\}_t, \{\lambda_{j,t}\}_{j,t}, \boldsymbol{\omega}, \{\gamma_{j,t}\}_{j,t}, \{\mu_j\}_j, \tau^2, \{\mathbf{n}^{(t)}\}_t\right) \\
&= \underbrace{p\left(\{\mathbf{n}^{(t)}\}_t \mid \{N^{(t)}\}_t, \{\lambda_{j,t}\}_{j,t}, \boldsymbol{\omega}, \boldsymbol{\gamma}, \{\mu_j\}_j, \tau^2\right)}_{\mathcal{L}_{complete}} \\
&\qquad * p\left(\{\lambda_{j,t}\}_{j,t} \mid \{N^{(t)}\}_t, \boldsymbol{\omega}, \boldsymbol{\gamma}, \{\mu_j\}_j, \tau^2\right) * \prod_{j=1}^{J} p(\mu_j) * p(\tau^2) * p(\boldsymbol{\gamma}) \\
&\qquad * p(\{N^{(t)}\}_t) * p(\boldsymbol{\omega}) \\
&= \prod_{t=1}^{T} \mathcal{L}_t(\boldsymbol{\lambda}_t, \boldsymbol{\omega}, N^{(t)}, \boldsymbol{\gamma}_t) \\
&\qquad * \prod_{j=1}^{J} p\left(\boldsymbol{\lambda}_j \mid \boldsymbol{\gamma}_j, \mu_j, \tau^2, \rho\right) * \prod_{j=1}^{J} p(\mu_j) * p(\tau^2) * (1/2)^{J*T} \\
&\qquad * \prod_{t=1}^{T} \frac{1}{N^{(t)}} * p(\boldsymbol{\omega})
\end{aligned}
$$

Our Gibbs sampler will iterate through the following steps:

1. Sample $\boldsymbol{\omega}$ using a Metropolis-Hastings scheme,

$$\left[\boldsymbol{\omega} \mid \{N^{(t)}\}_t, \{\lambda_{j,t}\}_{j,t}, \{\gamma_{j,t}\}_{j,t}, \{\mu_j\}_j, \tau^2, \{n_{\mathbf{k}}^{(t)}\}_{\mathbf{k}\neq\mathbf{o},t}\right] \propto \mathcal{L}_{complete} * p(\boldsymbol{\omega}).$$

   Propose new values by $sim \sim N_3(\mathbf{o}, \mathbb{I})$ and

$$\boldsymbol{\omega}^* = \boldsymbol{\omega}^{curr} + tun_\omega * \Sigma_{\boldsymbol{\omega}}^{prop} sim,$$

   where we take the square-root of the estimated covariance matrix for the MLE of the $\boldsymbol{\omega}$ from collapsing across years,

$$\Sigma_{\boldsymbol{\omega}}^{prop} = \widehat{Cov}\left(\widehat{\boldsymbol{\omega}}^{MLE}\right)^{1/2},$$

and $tun_{\boldsymbol{\omega}}$ can be adjusted to get an acceptance rate of about 25% (Roberts et al., 1994). This is a random walk by symmetry of the random noise, so the proposal distribution cancels and

$$R = \frac{\mathcal{L}_{complete}(\boldsymbol{\lambda}^{curr}, \boldsymbol{\omega}^*, \mathbf{N}^{curr}, \boldsymbol{\gamma}^{curr}) p(\boldsymbol{\omega}^*)}{\mathcal{L}_{complete}(\boldsymbol{\lambda}^{curr}, \boldsymbol{\omega}^{curr}, \mathbf{N}^{curr}, \boldsymbol{\gamma}^{curr}) p(\boldsymbol{\omega}^{curr})},$$

where we then accept the proposed value with probability $min(1, R)$.

2. Sample $\left[\{N^{(t)}\}_t \mid \boldsymbol{\omega}, \{\lambda_{j,t}\}_{j,t}, \{\gamma_{j,t}\}_{j,t}, \{\mu_j\}_j, \tau^2, \{n_{\mathbf{k}}^{(t)}\}_{\mathbf{k}\neq\mathbf{o},t}\right]$ which by properties of the SOUP is independent Negative Binomials (Meng and Zaslavsky, 2002):

$$\begin{aligned}
&\left[N^{(t)} \mid \{N^{(t')}\}_{t'\neq t}, \boldsymbol{\omega}, \{\lambda_{j,t}\}_{j,t}, \{\gamma_{j,t}\}_{j,t}, \{\mu_j\}_j, \tau^2, \{n_{\mathbf{k}}^{(t)}\}_{\mathbf{k}\neq\mathbf{o},t}\right] \\
&\quad \propto \mathcal{L}_t(\boldsymbol{\lambda}_t, \boldsymbol{\omega}, N^{(t)}, \boldsymbol{\gamma}_t) * \frac{1}{N^{(t)}} \\
&\quad \propto \frac{N^{(t)}!}{...n_{\mathbf{k}}^{(t)}!...(N^{(t)} - n^{(t)})!} \boldsymbol{\pi}_{\mathbf{o}}^{(t)}(\boldsymbol{\lambda}_t, \boldsymbol{\omega}, \boldsymbol{\gamma}_t)^{N^{(t)} - n^{(t)}} \prod_{\mathbf{k}\neq\mathbf{o}} \pi_{\mathbf{k}}^{(t)}(\boldsymbol{\lambda}_t, \omega, \gamma_t)^{n_{\mathbf{k}}^{(t)}} * \frac{1}{N^{(t)}} \\
&\quad \propto \frac{(N^{(t)} - 1)!}{(n^{(t)} - 1)!(N^{(t)} - n^{(t)})!} \boldsymbol{\pi}_{\mathbf{o}}^{(t)}(\boldsymbol{\lambda}_t, \boldsymbol{\omega}, \boldsymbol{\gamma}_t)^{N^{(t)} - n^{(t)}} (1 - \boldsymbol{\pi}_{\mathbf{o}}^{(t)}(\boldsymbol{\lambda}_t, \boldsymbol{\omega}, \boldsymbol{\gamma}_t))^{n^{(t)}} \\
&\quad \sim NegBin(n^{(t)}, 1 - \pi_{\mathbf{o}}^{(t)}(\boldsymbol{\lambda}_t, \boldsymbol{\omega}, \boldsymbol{\gamma}_t)).
\end{aligned}$$

3. Sample each $\boldsymbol{\lambda}_j$ using a Metropolis-Hastings scheme,

$$\begin{aligned}
&\left[\boldsymbol{\lambda}_j \;\middle|\; \{N^{(t)}\}_t, \boldsymbol{\omega}, \{\gamma_{j,t}\}_{j,t}, \{\mu_j\}_j, \tau^2, \{n_{\mathbf{k}}^{(t)}\}_{\mathbf{k}\neq\mathbf{o},t}\right] \\
&\qquad \propto \mathcal{L}_{complete}(\boldsymbol{\lambda}, \boldsymbol{\omega}, \mathbf{N}, \boldsymbol{\gamma}) * p\left(\boldsymbol{\lambda}_j \mid \boldsymbol{\gamma}_j, \mu_j, \tau^2\right).
\end{aligned}$$

Propose new values by $sim \sim N_T(0, \mathbb{I})$ and

$$\boldsymbol{\lambda}_j^* = \boldsymbol{\lambda}_j^{curr} + tun_\lambda * \Sigma_\lambda^{prop} sim,$$

where we take the square-root of the prior covariance matrix,

$$\Sigma_\lambda^{prop} = (\Sigma_j^{curr})^{1/2},$$

and as above tun_λ can be adjusted to get an acceptance rate of about 25% and the proposal distribution cancels so we construct R similarly to above and accept the proposed value with probability $min(1, R)$.

4. For $j = 1, ..., J$ and $t = 1, ..., T$, following logic similar to George and McCulloch (1993), we see that

$$\begin{aligned}\left[\gamma_{j,t} \;\middle|\; \{\lambda_{j,t}\}_{j,t}, \{N^{(t)}\}_t, \boldsymbol{\omega}, \{\gamma_{j',t'}\}_{-j,-t}, \{\mu_j\}_j, \tau^2, \{n_{\mathbf{k}}^{(t)}\}_{\mathbf{k}\neq\mathbf{o},t}\right] \\ \propto \mathcal{L}_t(\boldsymbol{\lambda}_t, \boldsymbol{\omega}, N^{(t)}, \boldsymbol{\gamma}_t) * p\left(\lambda_{j,t} \mid \gamma_{j,t}, \mu_j, \tau^2\right),\end{aligned}$$

so

$$\begin{aligned}P\left[\gamma_{j,t} = 1 \;\middle|\; \{\lambda_{j,t}\}_{j,t}, \{N^{(t)}\}_t, \boldsymbol{\omega}, \{\gamma_{j',t'}\}_{-j,-t}, \{\mu_j\}_j, \tau^2, \{n_{\mathbf{k}}^{(t)}\}_{\mathbf{k}\neq\mathbf{o},t}\right] \\ \propto \mathcal{L}_t(\boldsymbol{\lambda}_t, \boldsymbol{\omega}, N^{(t)}, \gamma_{j,t} = 1, \{\gamma_{j',t}\}_{-j}) * p\left(\lambda_{j,t} \mid \gamma_{j,t} = 1, \mu_j, \tau^2\right) \\ \equiv A\end{aligned}$$

and

$$\begin{aligned}P\left[\gamma_{j,t} = 0 \;\middle|\; \{\lambda_{j,t}\}_{j,t}, \{N^{(t)}\}_t, \boldsymbol{\omega}, \{\gamma_{j',t'}\}_{-j,-t}, \{\mu_j\}_j, \tau^2, \{n_{\mathbf{k}}^{(t)}\}_{\mathbf{k}\neq\mathbf{o},t}\right] \\ \propto \mathcal{L}_t(\boldsymbol{\lambda}_t, \boldsymbol{\omega}, N^{(t)}, \gamma_{j,t} = 0, \{\gamma_{j',t}\}_{-j}) * p\left(\lambda_{j,t} \mid \gamma_{j,t} = 0, \mu_j, \tau^2\right) \\ \equiv B,\end{aligned}$$

so we sample the $\gamma_{j,t}$ Bernoulli with probability $\frac{A}{A+B}$.

5. For $j = 1, ..., J$ sample

$$\begin{aligned}\left[\mu_j \;\middle|\; \{\beta_{j,t}\}_{j,t}, \{N^{(t)}\}_t, \boldsymbol{\omega}, \{\gamma_{j,t}\}_{j,t}, \{\mu_{j'}\}_{j'\neq j}, \{\tau_j^2\}_j, \{n_{\mathbf{k}}^{(t)}\}_{\mathbf{k}\neq\mathbf{o},t}\right] \\ \propto \prod_{t=1}^{T} \left[p_{N(\mu_{inactive}, \tau^2_{inactive})}\left(\beta_{j,t}\right)\right]^{1-\gamma_{j,t}} \left[p_{N(\mu_j, \tau_j^2)}\left(\beta_{j,t}\right)\right]^{\gamma_{j,t}} * p(\mu_j) \\ \propto \prod_{\substack{t=1 \\ \gamma_{j,t}=1}}^{T} p_{N(\mu_j, \tau_j^2)}\left(\beta_{j,t}\right) * p(\mu_j) \\ \sim N\left(\frac{\sigma_\mu^2 \sum_{\substack{t=1 \\ \gamma_{j,t}=1}}^{T} \beta_{j,t}}{\left[\sum_{t=1}^{T} \gamma_{j,t}\right] \sigma_\mu^2 + \tau_j^2}, \frac{\sigma_\mu^2 \tau_j^2}{\left[\sum_{t=1}^{T} \gamma_{j,t}\right] \sigma_\mu^2 + \tau_j^2}\right).\end{aligned}$$

6. Sample

$$\begin{aligned}
&\left[\tau^2 \;\middle|\; \{\beta_{j,t}\}_{j,t}, \{N^{(t)}\}_t, \boldsymbol{\omega}, \{\gamma_{j,t}\}_{j,t}, \{\mu_j\}_j, \{n^{(t)}_{\mathbf{k}}\}_{\mathbf{k}\neq\mathbf{o},t}\right] \\
&\quad\propto \prod_{j=1}^{J}\prod_{t=1}^{T}\left[p_{N(\mu_{inactive},\tau^2_{inactive}))}\left(\beta_{j,t}\right)\right]^{1-\gamma_{j,t}}\left[p_{N(\mu_j,\tau^2_j)}\left(\beta_{j,t}\right)\right]^{\gamma_{j,t}} * p(\tau^2) \\
&\quad\propto \prod_{j=1}^{J}\prod_{\substack{t=1\\ \gamma_{j,t}=1}}^{T}(\tau^2)^{-1/2}\exp(-1/2(\beta_{j,t}-\mu_j)^2/\tau^2) * (\tau^2)^{-a-1}\exp\left(\frac{-b}{\tau^2}\right) \\
&\quad\propto (\tau^2)^{-\sum_{j=1}^{J}\sum_{t=1}^{T}\gamma_{j,t}/2}\exp\left(-\frac{b+\sum_{j=1}^{J}\sum_{\substack{t=1\\ \gamma_{j,t}=1}}^{T}(\beta_{j,t}-\mu_j)^2/2}{\tau^2}\right) * (\tau^2)^{-a-1} \\
&\quad\sim IG\left(a+\left[\sum_{j=1}^{J}\sum_{t=1}^{T}\gamma_{j,t}\right]/2, b+\sum_{j=1}^{J}\sum_{\substack{t=1\\ \gamma_{j,t}=1}}^{T}(\beta_{j,t}-\mu_j)^2/2\right).
\end{aligned}$$

7. If we are fitting the zeros from missing data model, we also sample,

$$\left[\{n^{(t)}_{\mathbf{k}}\}_{\mathbf{k}\neq\mathbf{o},t} \;\middle|\; \{\beta_{j,t}\}_{j,t}, \{N^{(t)}\}_t, \boldsymbol{\omega}, \{\gamma_{j,t}\}_{j,t}, \{\mu_j\}_j, \tau^2, \{n^{(t)}_{\mathbf{k}}\}_{obs}\right]$$

If in year t, lists 3 and 4 are known to inactive, we use margins such as $n^{(t)}_{01{+}{+}0}$ or $n^{(t)}_{00{+}{+}0} = N^{(t)}$, and sample

$$\begin{aligned}
&(n^{(t)}_{01000}, n^{(t)}_{01010}, n^{(t)}_{01100}, n^{(t)}_{01110}) \\
&\qquad\sim Multi\left(n^{(t)}_{01{+}{+}0}, \frac{\left\{\pi^{(p,curr)}_{01000}, \pi^{(p,curr)}_{01010}, \pi^{(p,curr)}_{01100}, \pi^{(p,curr)}_{01110}\right\}}{\pi^{(p,curr)}_{01000}+\pi^{(p,curr)}_{01010}+\pi^{(p,curr)}_{01100}+\pi^{(p,curr)}_{01110}}\right)
\end{aligned}$$

The Gibbs sampling steps - Hierarchical AR1 model

The complete data likelihood $\mathcal{L}_{complete}$ is as in A.2.1 and the joint distribution is as in A.2.1, with the addition of the distribution for ρ, $p(\rho) = I(\rho \in (0,1))$, and ρ in the distribution

$p(\boldsymbol{\lambda}_j|\boldsymbol{\gamma}_j, \mu_j, \tau^2, \rho)$,

$$
\begin{aligned}
& p\left(\{N^{(t)}\}_t, \{\lambda_{j,t}\}_{j,t}, \boldsymbol{\omega}, \{\gamma_{j,t}\}_{j,t}, \{\mu_j\}_j, \tau^2, \rho, \{n_{\mathbf{k}}^{(t)}\}_{\mathbf{k}\neq\mathbf{o},t}\right) \\
&= \prod_{t=1}^{T} \mathcal{L}_t(\boldsymbol{\lambda}_t, \boldsymbol{\omega}, N^{(t)}, \boldsymbol{\gamma}_t) \\
&\quad * \prod_{j=1}^{J} p\left(\boldsymbol{\lambda}_j \mid \boldsymbol{\gamma}_j, \mu_j, \tau^2, \rho\right) * \prod_{j=1}^{J} p(\mu_j) * p(\tau^2) * I(\rho \in (0,1)) * (1/2)^{J*T} \\
&\quad * \prod_{t=1}^{T} \frac{1}{N^{(t)}} * p(\boldsymbol{\omega}).
\end{aligned}
$$

Our Gibbs sampler will iterate through the following steps:

1. Same as step 1 above.

2. Same as step 2 above.

3. Same as step 3 above, replacing $p\left(\boldsymbol{\lambda}_j \mid \boldsymbol{\gamma}_j, \mu_j, \tau^2\right)$ with$p\left(\boldsymbol{\lambda}_j \mid \boldsymbol{\gamma}_j, \mu_j, \tau^2, \rho\right)$.

4. Almost the same as step 4 above, for $j = 1, ..., J$ and $t = 1, ..., T$

$$
\begin{aligned}
&\left[\gamma_{j,t} \;\middle|\; \{\lambda_{j,t}\}_{j,t}, \{N^{(t)}\}_t, \boldsymbol{\omega}, \{\gamma_{j',t'}\}_{-j,-t}, \{\mu_j\}_j, \tau^2, \rho, \{n_{\mathbf{k}}^{(t)}\}_{\mathbf{k}\neq\mathbf{o},t}\right] \\
&\qquad \propto \mathcal{L}_t(\boldsymbol{\lambda}_t, \boldsymbol{\omega}, N^{(t)}, \boldsymbol{\gamma}_t) * p\left(\boldsymbol{\lambda}_j \mid \boldsymbol{\gamma}_j, \mu_j, \tau^2, \rho\right),
\end{aligned}
$$

so

$$
\begin{aligned}
&P\left[\gamma_{j,t} = 1 \;\middle|\; \{\lambda_{j,t}\}_{j,t}, \{N^{(t)}\}_t, \boldsymbol{\omega}, \{\gamma_{j',t'}\}_{-j,-t}, \{\mu_j\}_j, \tau^2, \rho, \{n_{\mathbf{k}}^{(t)}\}_{\mathbf{k}\neq\mathbf{o},t}\right] \\
&\qquad \propto \mathcal{L}_t(\boldsymbol{\lambda}_t, \boldsymbol{\omega}, N^{(t)}, \gamma_{j,t} = 1, \{\gamma_{j',t}\}_{-j}) * p\left(\boldsymbol{\lambda}_j \mid \gamma_{j,t} = 1, \{\gamma_{j',t}\}_{-j}, \mu_j, \tau^2, \rho\right) \\
&\qquad \equiv A
\end{aligned}
$$

and

$$
\begin{aligned}
&P\left[\gamma_{j,t} = 0 \;\middle|\; \{\lambda_{j,t}\}_{j,t}, \{N^{(t)}\}_t, \boldsymbol{\omega}, \{\gamma_{j',t'}\}_{-j,-t}, \{\mu_j\}_j, \tau^2, \rho, \{n_{\mathbf{k}}^{(t)}\}_{\mathbf{k}\neq\mathbf{o},t}\right] \\
&\qquad \propto \mathcal{L}_t(\boldsymbol{\lambda}_t, \boldsymbol{\omega}, N^{(t)}, \gamma_{j,t} = 0, \{\gamma_{j',t}\}_{-j}) * p\left(\boldsymbol{\lambda}_j \mid \gamma_{j,t} = 0, \{\gamma_{j',t}\}_{-j}, \mu_j, \tau^2, \rho\right) \\
&\qquad \equiv B,
\end{aligned}
$$

so we sample the $\gamma_{j,t}$ Bernoulli with probability $\frac{A}{A+B}$.

5. For $j = 1, ..., J$, sample

$$\begin{aligned}
&\left[\mu_j \;\middle|\; \{\lambda_{j,t}\}_{j,t}, \{N^{(t)}\}_t, \boldsymbol{\omega}, \{\gamma_{j,t}\}_{j,t}, \{\mu_{j'}\}_{-j}, \tau^2, \rho, \{n_{\mathbf{k}}^{(t)}\}_{\mathbf{k}\neq\mathbf{o},t}\right] \\
&\quad \propto p\left(\boldsymbol{\lambda}_j \mid \boldsymbol{\gamma}_j, \mu_j, \tau^2, \rho\right) * p(\mu_j) \\
&\quad \propto \exp\left(-\frac{1}{2}\mu_j^2\frac{1}{\sigma_\mu^2} - \frac{1}{2}(\boldsymbol{\lambda}_j - (1-\boldsymbol{\gamma}_j)\mu_{inactive} - \boldsymbol{\gamma}_j\mu_j)'\Sigma_j^{-1}(\boldsymbol{\lambda}_j - (1-\boldsymbol{\gamma}_j)\mu_{inactive} - \boldsymbol{\gamma}_j\mu_j)\right) \\
&\quad \propto \exp\left[-\frac{1}{2}\left(\mu_j^2\frac{1}{\sigma_\mu^2} + (\boldsymbol{\lambda}_j - (1-\boldsymbol{\gamma}_j)\mu_{inactive} - \boldsymbol{\gamma}_j\mu_j)'\Sigma_j^{-1}(\boldsymbol{\lambda}_j - (1-\boldsymbol{\gamma}_j)\mu_{inactive} - \boldsymbol{\gamma}_j\mu_j)\right)\right] \\
&\quad \propto \exp\left[-\frac{1}{2}\left(\mu_j^2\underbrace{\left(\boldsymbol{\gamma}_j'\Sigma_j^{-1}\boldsymbol{\gamma}_j + \frac{1}{\sigma_\mu^2}\right)}_{a} - \mu_j * \underbrace{2\left(\boldsymbol{\lambda}_j'\Sigma_j^{-1}\boldsymbol{\gamma}_j + \mu_{inactive}(1-\boldsymbol{\gamma}_j)'\Sigma_j^{-1}\boldsymbol{\gamma}_j\right)}_{b}\right)\right] \\
&\quad \propto \exp\left[-\frac{1}{2}a(\mu_j - b/(2a))^2\right] \\
&\quad \sim N\left(b/(2a), 1/a\right)
\end{aligned}$$

6. Sample τ^2 using a Metropolis-Hastings scheme,

$$\begin{aligned}
&\left[\tau^2 \;\middle|\; \{\lambda_{j,t}\}_{j,t}, \{N^{(t)}\}_t, \boldsymbol{\omega}, \{\gamma_{j,t}\}_{j,t}, \{\mu_j\}_j, \rho, \{n_{\mathbf{k}}^{(t)}\}_{\mathbf{k}\neq\mathbf{o},t}\right] \\
&\quad \propto \prod_{j=1}^{J} p\left(\boldsymbol{\lambda}_j \mid \boldsymbol{\gamma}_j, \mu_j, \tau^2, \rho\right) * p(\tau^2) \\
&\quad \propto \exp\left(-\frac{1}{2}\sum_{j=1}^{J}(\boldsymbol{\lambda}_j - (1-\boldsymbol{\gamma}_j)\mu_{inactive} - \boldsymbol{\gamma}_j\mu_j)'\Sigma_j^{-1}(\boldsymbol{\lambda}_j - (1-\boldsymbol{\gamma}_j)\mu_{inactive} - \boldsymbol{\gamma}_j\mu_j)\right) \\
&\qquad * \prod_{j=1}^{J}|2\pi\Sigma_j|^{-1/2} * (\tau^2)^{-a-1}\exp\left(\frac{-b}{\tau^2}\right).
\end{aligned}$$

Propose new values by $sim \sim N_1(0,1)$ and

$$\log(\tau^{2,*}) = \log(\tau^{2,curr}) + tun_\tau * sim$$

$$\tau^{2,*} = \exp(\log(\tau^{2,curr}) + tun_\tau * sim)$$

and as above tun_τ can be adjusted to get an acceptance rate of about 45% and the proposal distribution cancels so we construct R similarly to above and accept the proposed value with probability $min(1, R)$ (Roberts et al., 1994).

7. Sample ρ using a Metropolis-Hastings scheme,

$$\left[\rho \;\middle|\; \{\lambda_{j,t}\}_{j,t}, \{N^{(t)}\}_t, \boldsymbol{\omega}, \{\gamma_{j,t}\}_{j,t}, \{\mu_j\}_j, \tau^2, \{n_{\mathbf{k}}^{(t)}\}_{\mathbf{k}\neq\mathbf{o},t}\right]$$
$$\propto \prod_{j=1}^{J} p\left(\boldsymbol{\lambda}_j \mid \boldsymbol{\gamma}_j, \mu_j, \tau^2, \rho\right) * p(\rho)$$
$$\propto \exp\left(-\frac{1}{2}\sum_{j=1}^{J}(\boldsymbol{\lambda}_j - (1-\boldsymbol{\gamma}_j)\mu_{inactive} - \boldsymbol{\gamma}_j\mu_j)'\Sigma_j^{-1}(\boldsymbol{\lambda}_j - (1-\boldsymbol{\gamma}_j)\mu_{inactive} - \boldsymbol{\gamma}_j\mu_j)\right)$$
$$* I(\rho \in (0,1)).$$

Propose new values by $sim \sim N_1(0,1)$ and

$$\rho^* = \rho^{curr} + tun_\rho * sim$$

and as above tun_ρ can be adjusted to get an acceptance rate of about 45% and the proposal distribution cancels so we construct R similarly to above and accept the proposed value with probability $min(1, R)$.

SOUP prior uninformativeness

The *single observation unbiased prior* (SOUP) was developed by Meng and Zaslavsky (2002) for $\pi_{\mathbf{o}}$ (probability of being unrecorded) known, and extended for capture-recapture in Stuart and Zaslavsky (2005). With the SOUP, the posterior mean of the total number of events is unbiased,

$$E[N^*|N, \pi_{\mathbf{o}}] = E_{n\sim Bin(N,1-\pi_{\mathbf{o}})}\left[E_{N^*\sim NegBin(n,1-\pi_{\mathbf{o}})}\left(N^*|n, \pi_{\mathbf{o}}\right)\middle| N, \pi_{\mathbf{o}}\right] = N.$$

Here we extend the work in Stuart and Zaslavsky (2005) to our situation, with multiple years and more than two lists. We take the yearly population total priors to be independent $\pi(N_t) \propto 1/N_t$. We show that the SOUP is uninformative for the cell probabilities by integrating out $\{N^{(t)}\}_t$ from the posterior to show that we get the same inference as on

the observed cells only. The posterior is

$$
\begin{aligned}
& p\left(\{N^{(t)}\}_t, \{\lambda_{j,t}\}_{j,t}, \boldsymbol{\omega}, \{\gamma_{j,t}\}_{j,t}, \{\mu_j\}_j, \tau^2, \rho, \mid \{n_{\mathbf{k}}^{(t)}\}_{\mathbf{k}\neq\mathbf{o},t}\right) \\
& \propto p\left(\{N^{(t)}\}_t, \{\lambda_{j,t}\}_{j,t}, \boldsymbol{\omega}, \{\gamma_{j,t}\}_{j,t}, \{\mu_j\}_j, \tau^2, \rho, \{n_{\mathbf{k}}^{(t)}\}_{\mathbf{k}\neq\mathbf{o},t}\right) \\
& = p\left(\{N^{(t)}\}_t, \{\lambda_{j,t}\}_{j,t}, \boldsymbol{\omega}, \{\gamma_{j,t}\}_{j,t}, \{\mu_j\}_j, \tau^2, \rho, \{\mathbf{n}^{(t)}\}_t\right) \\
& = \prod_{t=1}^{T} \binom{N^{(t)}}{\{n_{\mathbf{k}}^{(t)}\}_{\mathbf{k}}} \prod_{\mathbf{k}} \boldsymbol{\pi}_{\mathbf{k}}^{(t)}(\boldsymbol{\lambda}_t, \boldsymbol{\omega}, \boldsymbol{\gamma}_t)^{n_{\mathbf{k}}^{(t)}} \\
& \quad * \prod_{j=1}^{J} p\left(\boldsymbol{\lambda}_j \mid \boldsymbol{\gamma}_j, \mu_j, \tau^2, \rho\right) * \prod_{j=1}^{J} p(\mu_j) * p(\tau^2) * 1 * (1/2)^{J*T} \\
& \quad * \prod_{t=1}^{T} \frac{1}{N^{(t)}} * p(\omega_{NGO}) p(\omega_{govt}) p(\omega_{mix}).
\end{aligned}
$$

Integrating out $\{N^{(t)}\}_t$,

$$
\begin{aligned}
& p\left(\{\lambda_{j,t}\}_{j,t}, \boldsymbol{\omega}, \{\gamma_{j,t}\}_{j,t}, \{\mu_j\}_j, \tau^2, \rho, \mid \{n_{\mathbf{k}}^{(t)}\}_{\mathbf{k}\neq\mathbf{o},t}\right) \\
& \propto \sum_{N^{(1)}=n^{(1)}}^{\infty} \ldots \sum_{N^{(T)}=n^{(T)}}^{\infty} p\left(\{N^{(t)}\}_t, \{\lambda_{j,t}\}_{j,t}, \boldsymbol{\omega}, \{\gamma_{j,t}\}_{j,t}, \{\mu_j\}_j, \tau^2, \rho, \{n_{\mathbf{k}}^{(t)}\}_{\mathbf{k}\neq\mathbf{o},t}\right) \\
& \text{letting } \boldsymbol{\pi}_{\mathbf{k}}^{(t)}(\boldsymbol{\lambda}_t, \boldsymbol{\omega}, \boldsymbol{\gamma}_t) = \boldsymbol{\pi}_{\mathbf{k}}^{(t)}, \\
& = \sum_{N^{(1)}=n^{(1)}}^{\infty} \ldots \sum_{N^{(T)}=n^{(T)}}^{\infty} \prod_{t=1}^{T} \binom{N^{(t)}}{\{n_{\mathbf{k}}^{(t)}\}_{\mathbf{k}}} \prod_{\mathbf{k}} \boldsymbol{\pi}_{\mathbf{k}}^{(t)^{n_{\mathbf{k}}^{(t)}}} * \frac{1}{N^{(t)}} \\
& \qquad * \left\{\text{priors for log-linear parameters and hyperparameters}\right\} \\
& = \prod_{t=1}^{T} \sum_{N^{(t)}=n^{(t)}}^{\infty} \binom{N^{(t)}}{\{n_{\mathbf{k}}^{(t)}\}_{\mathbf{k}}} \prod_{\mathbf{k}} \boldsymbol{\pi}_{\mathbf{k}}^{(t)^{n_{\mathbf{k}}^{(t)}}} * \frac{1}{N^{(t)}} \\
& \qquad * \left\{\text{priors for log-linear parameters and hyperparameters}\right\} \\
& = \prod_{t=1}^{T} \prod_{\mathbf{k}\neq\mathbf{o}} \boldsymbol{\pi}_{\mathbf{k}}^{(t)^{n_{\mathbf{k}}^{(t)}}} \frac{1}{\prod_{\mathbf{k}\neq\mathbf{o}} n_{\mathbf{k}}^{(t)}} \sum_{N^{(t)}=n^{(t)}}^{\infty} \frac{(N^{(t)}-1)!}{(N^{(t)}-n^{(t)})!} \boldsymbol{\pi}_{\mathbf{o}}^{(t)^{n_{\mathbf{o}}^{(t)}}} \\
& \qquad * \left\{\text{priors for log-linear parameters and hyperparameters}\right\} \\
& = \prod_{t=1}^{T} \prod_{\mathbf{k}\neq\mathbf{o}} \boldsymbol{\pi}_{\mathbf{k}}^{(t)^{n_{\mathbf{k}}^{(t)}}} \frac{1}{\prod_{\mathbf{k}\neq\mathbf{o}} n_{\mathbf{k}}^{(t)}} \underbrace{\sum_{N^{(t)}=n^{(t)}}^{\infty} \frac{(N^{(t)}-1)!}{(N^{(t)}-n^{(t)})!} \boldsymbol{\pi}_{\mathbf{o}}^{(t)^{n_{\mathbf{o}}^{(t)}}} \times \frac{(1-\boldsymbol{\pi}_{\mathbf{o}}^{(t)})^{n^{(t)}}}{(n^{(t)}-1)!}}_{NegBin \text{ so}=1} \frac{(n^{(t)}-1)!}{(1-\boldsymbol{\pi}_{\mathbf{o}}^{(t)})^{n^{(t)}}} \\
& \qquad * \left\{\text{priors for log-linear parameters and hyperparameters}\right\}
\end{aligned}
$$

which is an independent multinomial on observed cells for each year. Thus, the SOUP prior is uninformative for the cell probabilities.

A.2.2 Data Descriptives

Here we include some summaries of the Casanare data and details about the matching algorithm.

Table A.5: Distribution of the number of times a record appears on the lists, i.e. the number of captures.

Killings	
Number of captures	Frequency
1	1871
2	571
3	157
4	14
5	6
6	0

Table A.6: We display for each list the number of recorded killings between 1998 and 2007, the type of list (government or non-government organization) and abbreviation used in the paper.

Organization name	Recorded Killings	Type	Abbreviation
National Institute of Forensic Medicine Deaths	1874	govt	IMLM
National Police - DIJIN	825	govt	PN0
Human Rights Observatory of the Vice Presidency	501	govt	VP
Colombian Commission of Jurists	160	NGO	CCJ
CINEP	138	NGO	CIN
Colombia-Europe-US Coordination	72	NGO	CCE

Matching rules

From Guberek et al. (2010), the following rules were used to match the data by the Human Rights Data Analysis Group (HRDAG). HRDAG grouped together multiple records on the same victim into a "match group." When the records were not exact matches on violation type, contradictions were resolved by the rules:

- If at least one record in the match group was a killing, the group was determined to be a killing.
- If at least one record was a disappearance and there was no killing record in the group, the group was considered a disappearance.
- Records of detentions, hostages and extortive kidnappings were only kept if they were matched with a record of killing or disappearance, the rest were dropped from the data.

The data were matched by two human matchers, who showed a high rate of agreement, with 280 pairs of records matched by the first matcher but not the second, 149 pairs matched by the second matcher but not the first, and 4,389 pairs of records matched by both of them (Guberek et al., 2010).

A.2.3 Extra Posterior Predictive Checks for the Casanare Data

Figures A.14, A.15, A.16 show graphical/visual assessments of the fit of models H-ZS, AR1-ZS and H-ZM to the Casanare data. Each box of gray-scale rectangles represents a table of cell counts, with years as the columns and rows as the capture histories. The top row represents a capture history $\mathbf{k} = 000000$, and the bottom row $\mathbf{k} = 111111$, with capture histories in between listed in increasing order as binary numbers. The grayscale represents the cell count, with darker indicating a higher count and white indicating a

zero cell. We visually inspect the boxes for the similarity of simulated data to the true cell counts from Casanare.

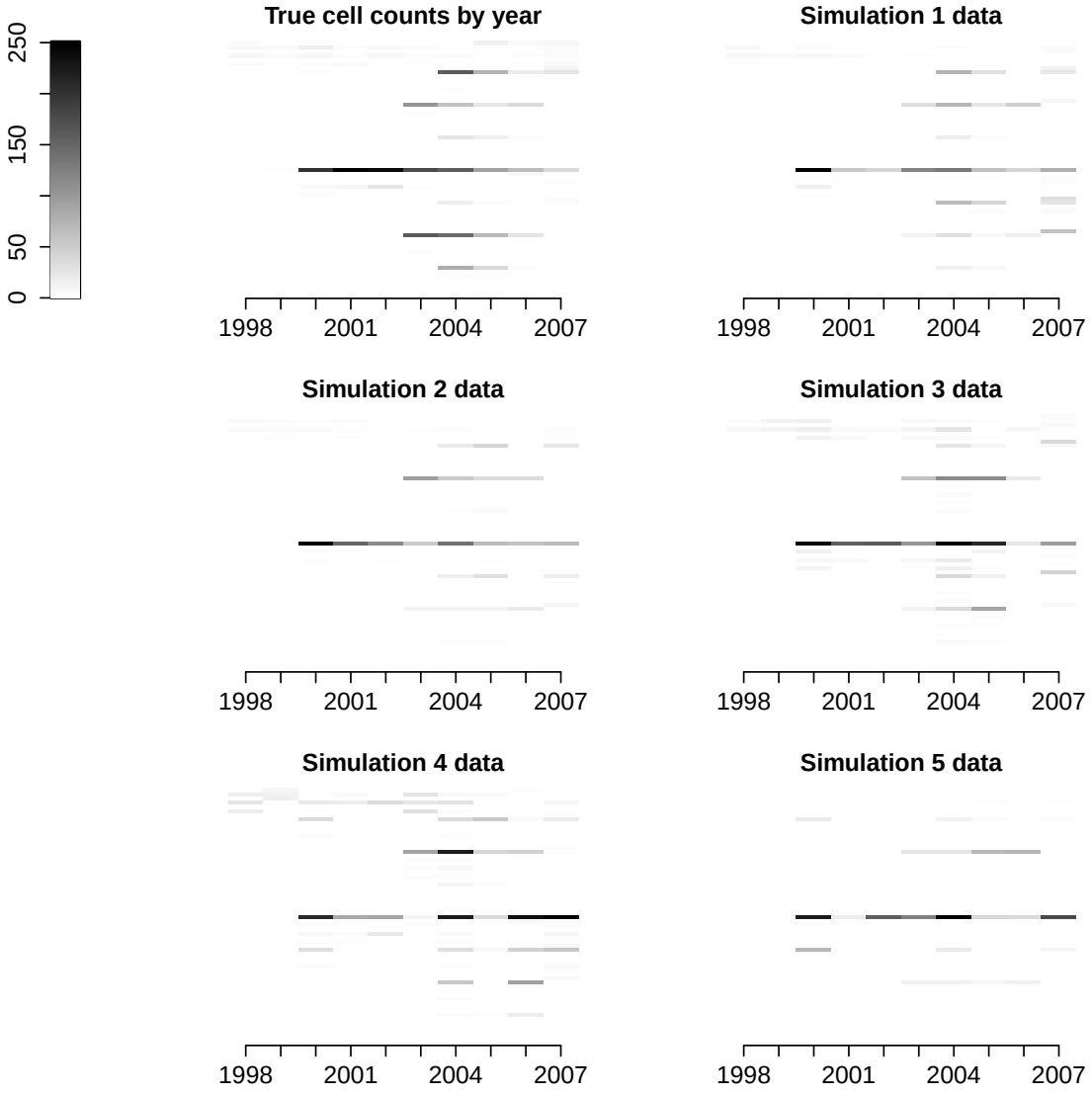


Figure A.14: Posterior predictive check for the H-ZS model: graphical/visual assessments of the fit of models H-ZS, AR1-ZS and H-ZM to the Casanare data. Each box of gray-scale rectangles represents a table of cell counts, with years as the columns and rows as the capture histories. The top row represents a capture history $\mathbf{k} = 000000$, and the bottom row $\mathbf{k} = 111111$, with capture histories in between listed in increasing order as binary numbers. The grayscale represents the cell count, with darker indicating a higher count and white indicating a zero cell. We visually inspect the boxes for the similarity of simulated data to the true cell counts from Casanare.

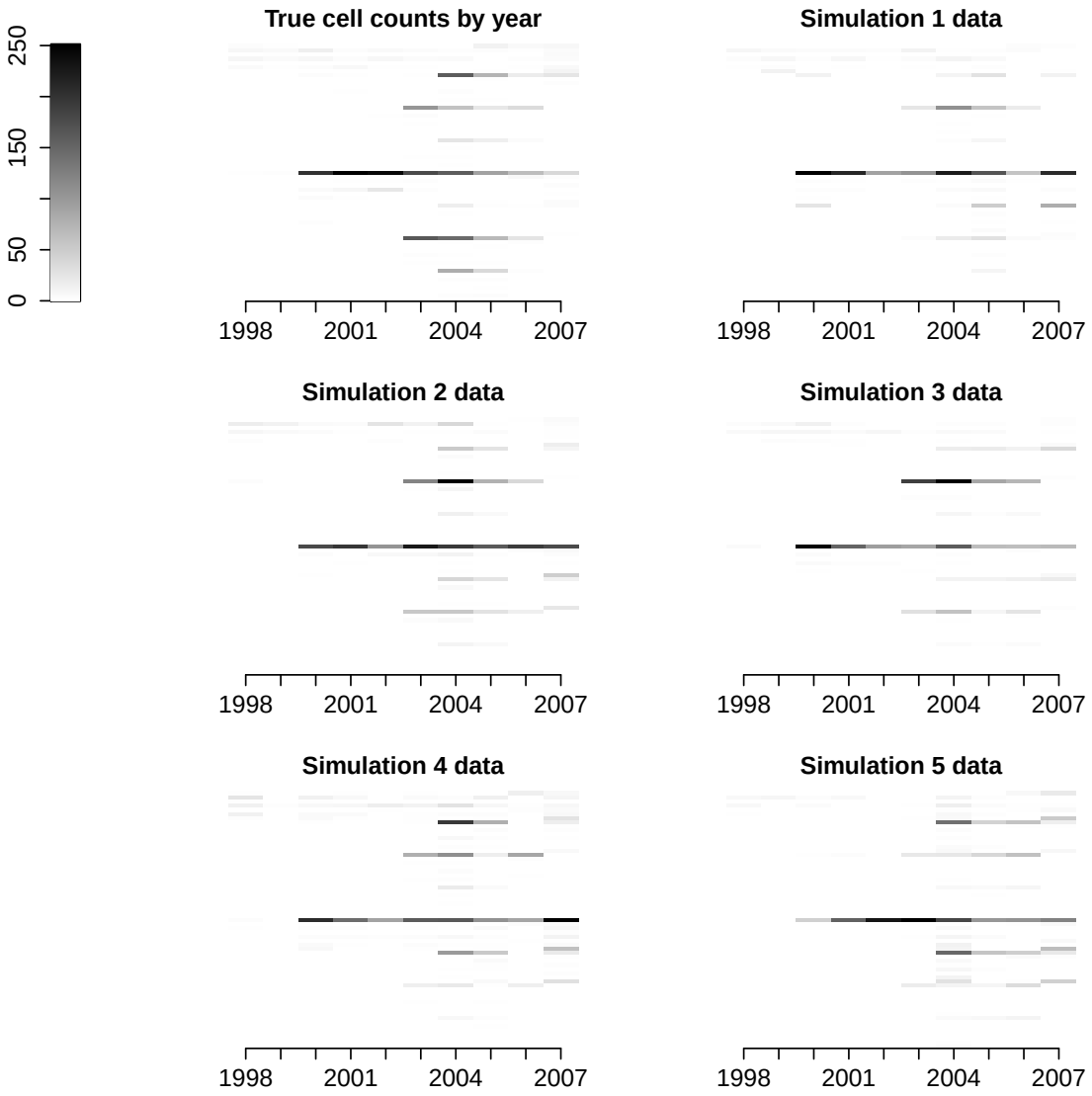


Figure A.15: Posterior predictive check for the AR1-ZS model: graphical/visual assessments of the fit of models H-ZS, AR1-ZS and H-ZM to the Casanare data. Each box of gray-scale rectangles represents a table of cell counts, with years as the columns and rows as the capture histories. The top row represents a capture history $\mathbf{k} = 000000$, and the bottom row $\mathbf{k} = 111111$, with capture histories in between listed in increasing order as binary numbers. The grayscale represents the cell count, with darker indicating a higher count and white indicating a zero cell. We visually inspect the boxes for the similarity of simulated data to the true cell counts from Casanare.

A.2.4 Extra Simulation Results

The fraction of the simulation where estimates $\widehat{N}^{(t)}$'s are higher than the population of Casanare, when generating data from the H-ZS model:

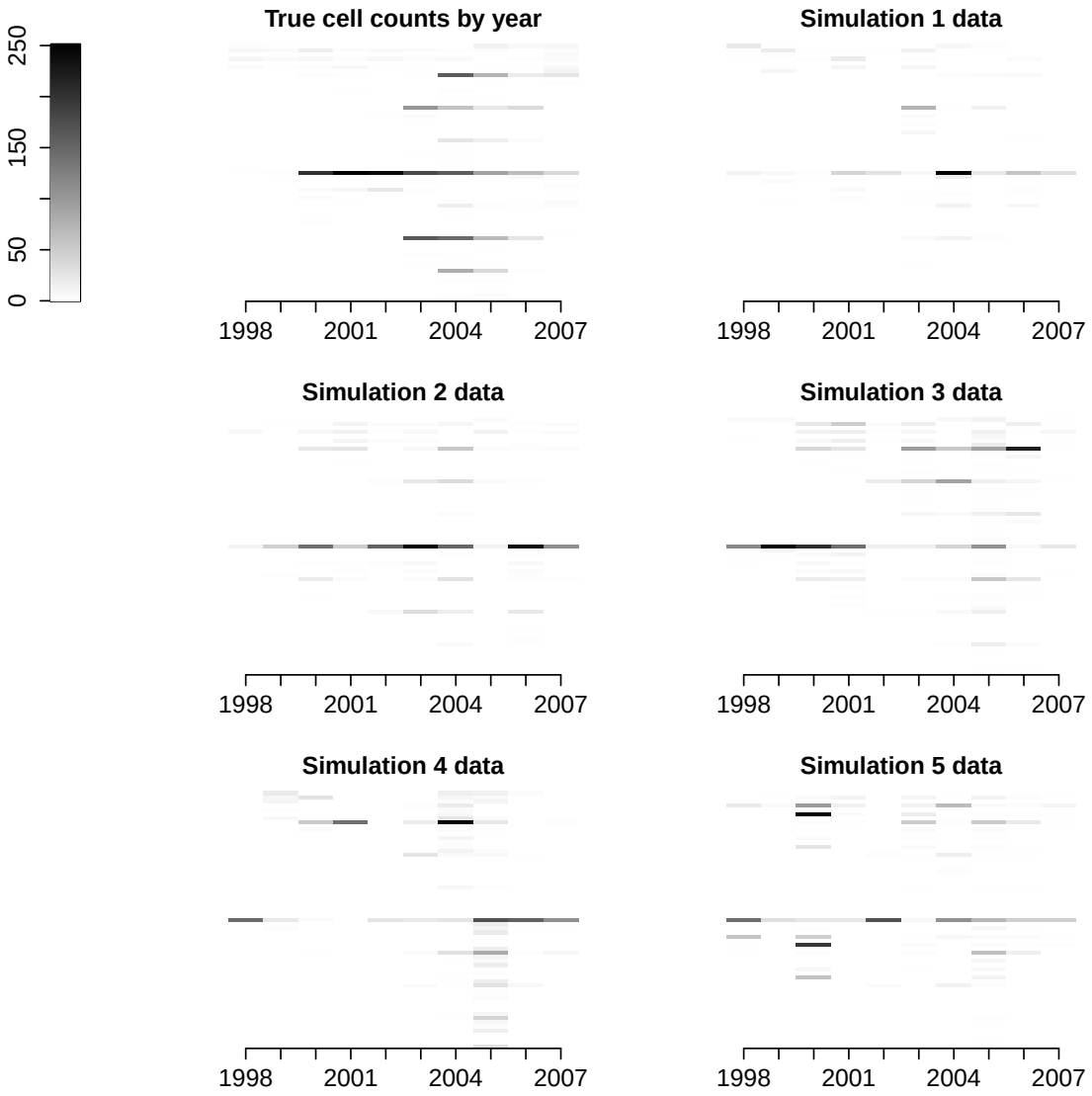


Figure A.16: Posterior predictive check for the H-ZM model: graphical/visual assessments of the fit of models H-ZS, AR1-ZS and H-ZM to the Casanare data. Each box of gray-scale rectangles represents a table of cell counts, with years as the columns and rows as the capture histories. The top row represents a capture history $\mathbf{k} = 000000$, and the bottom row $\mathbf{k} = 111111$, with capture histories in between listed in increasing order as binary numbers. The grayscale represents the cell count, with darker indicating a higher count and white indicating a zero cell. We visually inspect the boxes for the similarity of simulated data to the true cell counts from Casanare.

	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Separate each year	0.19	0.2	0.24	0.37	0.60	0.01	0	0	0.05	0
U-ZS	0.19	0.2	0.01	0.03	0.25	0.00	0	0	0.00	0

H-ZS	0.00	0.0	0.00	0.00	0.00	0.00	0	0	0.00	0
AR1-ZS	0.00	0.0	0.00	0.00	0.00	0.00	0	0	0.00	0
U-ZM	0.19	0.2	0.01	0.03	0.25	0.00	0	0	0.00	0
H-ZM	0.00	0.0	0.00	0.00	0.00	0.00	0	0	0.00	0

The fraction of the simulation where estimates $\widehat{N}^{(t)}$'s are higher than the population of Casanare, when generating data from the AR1-ZS model:

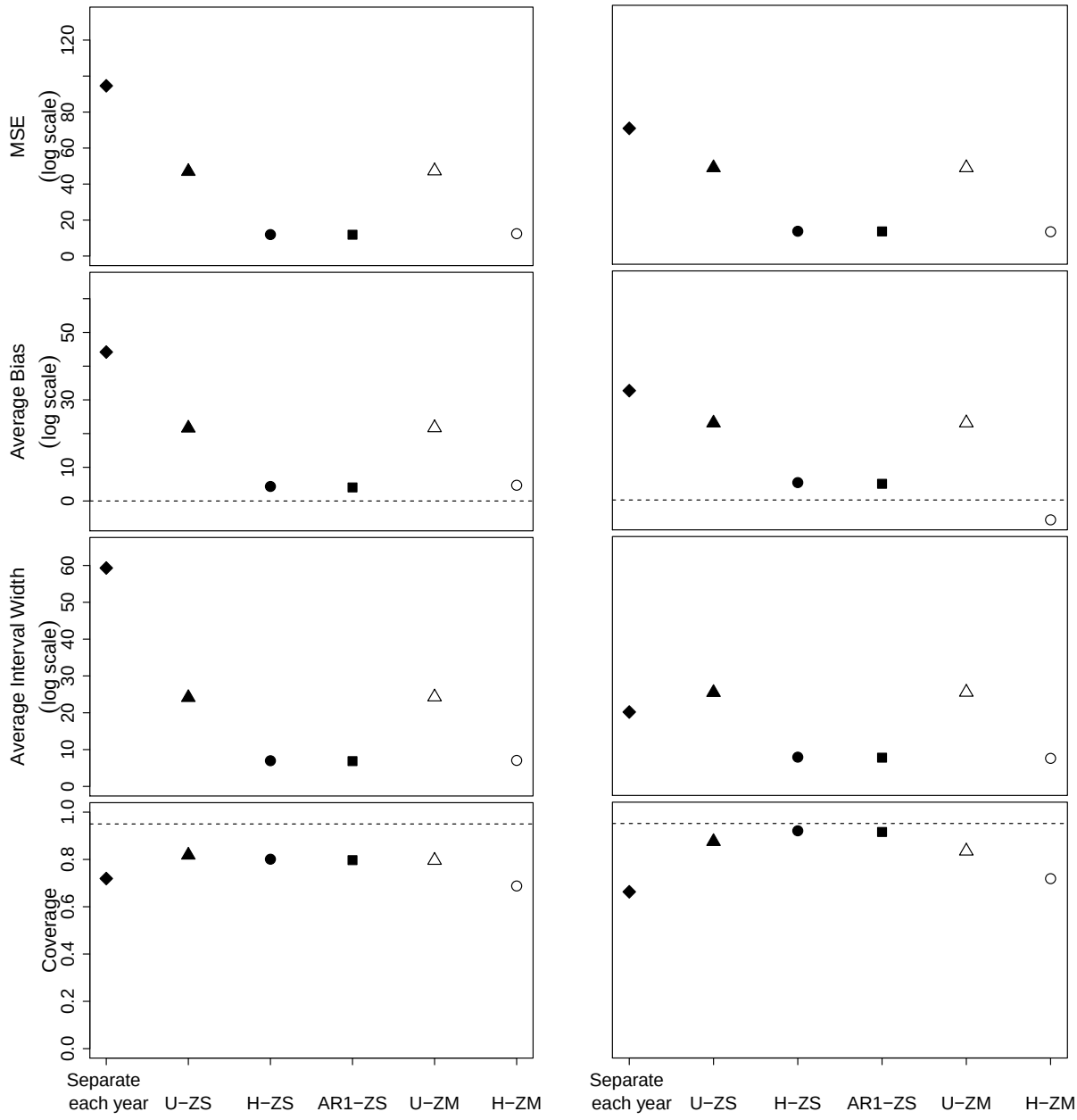
	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Separate each year	0.18	0.2	0.17	0.27	0.69	0	0	0	0.08	0
U-ZS	0.17	0.2	0.00	0.03	0.18	0	0	0	0.00	0
H-ZS	0.00	0.0	0.00	0.00	0.00	0	0	0	0.00	0
AR1-ZS	0.00	0.0	0.00	0.00	0.00	0	0	0	0.00	0
U-ZM	0.17	0.2	0.00	0.03	0.18	0	0	0	0.00	0
H-ZM	0.00	0.0	0.00	0.00	0.00	0	0	0	0.00	0

The fraction of the simulation where estimates $\widehat{N}^{(t)}$'s are higher than the population of Casanare, when generating data from the H-ZM model:

	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
Separate each year	0	0	0	0.05	0.41	0	0	0	0.01	0
U-ZS	0	0	0	0.00	0.00	0	0	0	0.00	0
H-ZS	0	0	0	0.00	0.00	0	0	0	0.00	0
AR1-ZS	0	0	0	0.00	0.00	0	0	0	0.00	0
U-ZM	0	0	0	0.00	0.00	0	0	0	0.00	0
H-ZM	0	0	0	0.00	0.00	0	0	0	0.00	0

A.2.5 Acknowledgments

The authors thank HRDAG for the inspiration for this work, their cleaned and matched data, and subject-matter knowledge. In particular, we thank Megan Price and Patrick Ball for guidance in the use of capture-recapture methods for human rights data. The authors also thank Professors Peter van der Heijden and Alan Agresti for input and expertise in capture-recapture methods. This material is based upon work supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE-1144152.



(a) AR1-ZS, all years

(b) AR1-ZS, 1998-1999

Figure A.17: Results from simulations, generating data from the AR1-ZS model, with $\rho = 0.5$, using μ_j, τ^2, ω, $N^{(t)}$, and $\gamma_{j,t}$ from posterior means of Casanare data. We show results from all years, and from 1998 and 1999 alone. We do 100 simulations from the AR1-ZS model.

A.3 The Millennium Villages Project: A protocol for the final evaluation

A.4 Timeline of Key Interventions

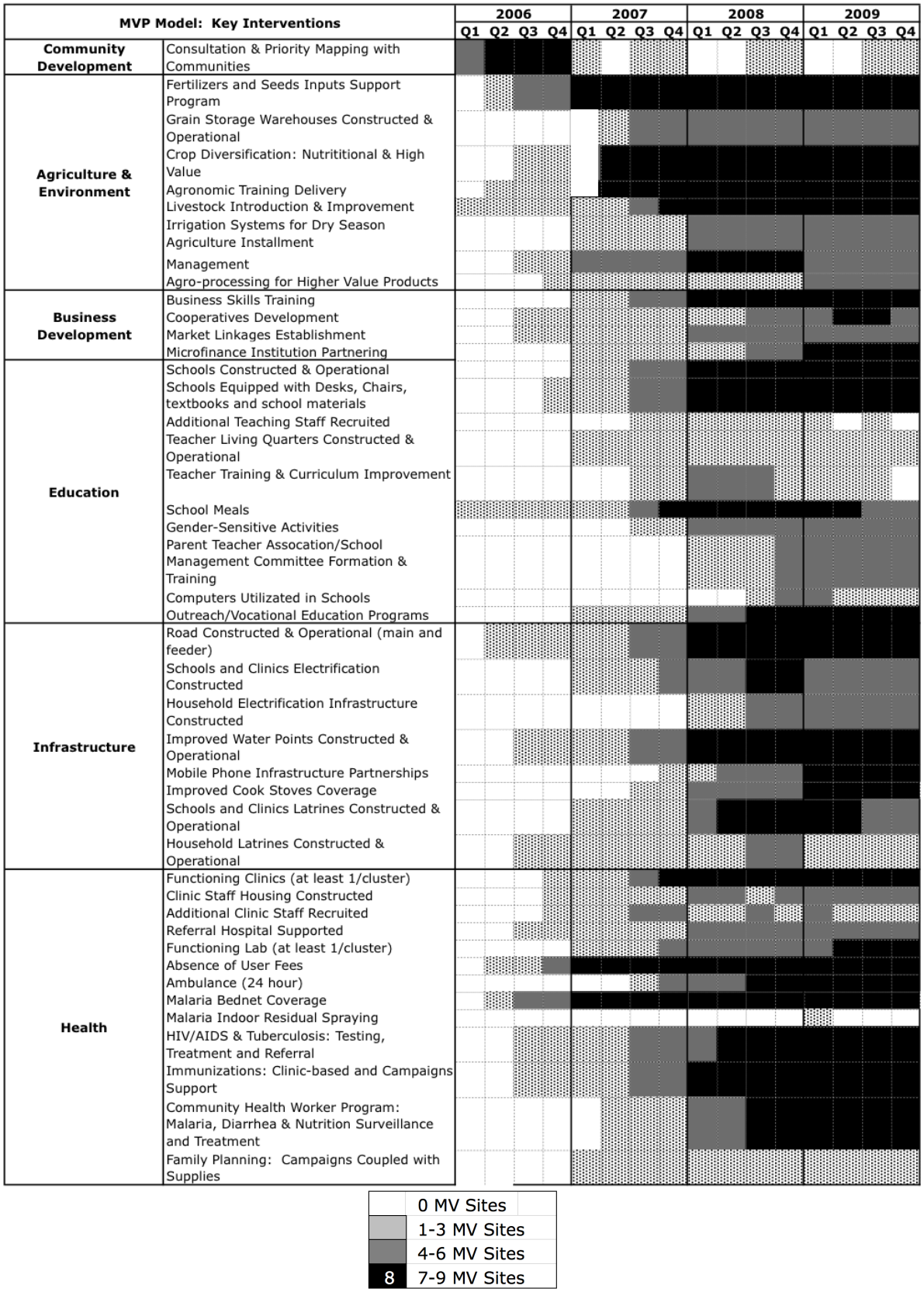


Figure A.18: Timeline of Key Interventions.

A.5 MDG Targets per MVP village

Data sources used for Table A.7 include:

(WB) World Bank PovCal: http://iresearch.worldbank.org/PovcalNet/index.htm?3

(WN) WHO NLIS: http://apps.who.int/nutrition/landscape/search.aspx?dm=52&countries=

(W) WHO: http://apps.who.int/ghodata/?theme=country

(D) DHS: http://www.measuredhs.com/data/available-datasets.cfm or http://www.statcompiler.com/

(U) UNSTATS: http://mdgs.un.org/unsd/mdg/Default.aspx

(M) MICS: http://www.unicef.org/statistics/index_24302.html

Table A.7: MVP 1990 National and Rural Reference Data Used to Set 2015 Targets

#	Indicator	Koraro, Ethiopia	Bonsasso, Ghana	Sauri, Kenya	Mwandama, Malawi	Tiby, Mali	Pampaida, Nigeria	Mayange, Rwanda	Potou, Senegal	Mbola, Tanzania	Ruhiira, Uganda
1.1 (WB)	Proportion of population below 1.25 USD (PPP 2005) per day	60.5 (1995)	51.1 (1992)	38.4 (1992)	83.1% (1998)	86.1% (1993)	61.9% (1992)	74.6% (2000)	65.8% (1991)	72.9% (1992)	70.0% (1992)
1.2 (WB)	Poverty Gap ratio	21.2 (1995)	18.3 (1992)	15.4 (1992)	46.0 (1998)	53.1 (1994)	31.1 (1992)	36.9 (2000)	34.3 (1991)	29.7 (1992)	30.3 (1992)
1.8 (WN)	Underweight among children under 5 years old	43.5% (2000)	22.6% (1999)	21.2% (1993)	25.6% (1992)	33.9% (2001)	38.2% (1990)	24.7% (1992)	27.3% (1993)	26.0% (1992)	20.0% (2001)
2.1 (D)	Net attendance ratio in primary education	24.3% (2000)	70.3 % (1993)	84.1% (1993)	49.6% (1992)	15.9% (1995)	46.6% (1990)	46.7% (1992)	16.0% (1992)	24.6 % (1991)	62.2% (1995)
2.2 (U)	Proportion of pupils starting grade 1 who reach last grade of primary education	23% (1994)	63 % (1991)	73% (2003)	24% (1990)	65% (1999)	73% (2002)	41% (1990)	69% (1990)	58% (1992)	38% (2000)
3.1 (U)	Gender parity in primary education	0.66 (1991)	0.86 (1991)	0.97 (1991)	0.87 (1991)	0.61 (1991)	0.81 (1991)	0.99 (1991)	0.73 (1991)	0.98 (1991)	0.81 (1991)
4.1 (D)	Under-5 mortality rate (per 1000 births)	192 (2000)	149 (1993)	96 (1993)	244 (1992)	273 (1995)	208 (1990)	163 (1992)	184 (1992)	152 (1991)	159 (1995)
4.2 (D)	Infant mortality rate (per 1000 births)	115 (2000)	82 (1993)	65 (1993)	138 (1992)	145 (1995)	96 (1990)	90 (1992)	87 (1992)	97 (1991)	88 (1995)
4.3 (UNI/W)	Measles immunization rate of 1 year-old children	38% (1990)	61% (1990)	78% (1990)	81% (1990)	43% (1990)	54% (1990)	83% (1990)	51% (1990)	80% (1990)	52% (1990)
5.2 (M/D)	Skilled birth attendance	5.6% (2000)	43.8% (1993)	45.4% (1993)	54.8% (1992)	40.0% (1995)	30.8% (1990)	25.8% (1992)	47.2% (1993)	43.9% (1992)	37.8% (1995)
5.3 (D)	Modern contraception use	3.3% (2000)	7.4% (1993)	25.4% (1993)	6.0% (1992)	1.9% (1995)	1.9% (1990)	12.6 % (1992)	1.4% (1992)	4.4% (1991)	5.1% (1995)
5.5 (D)	Antenatal care coverage	10.4% (2000)	58.9% (1993)	63.7% (1993)	62.6 % (1992)	25.8% (1995)	51.5% (1990)	12.0% (1992)	13.3% (1992)	69.3% (1991)	47.2% (1995)
6.7 (U)	Children under 5 sleeping under insecticide-bed nets	1.5% (2005)	3.5% (2003)	2.9% (2000)	2.8% (2000)	27.1% (2006)	1.2% (2003)	5.0% (2000)	1.7% (2000)	2.1% (1999)	0.2% (2001)
7.8 (U)	Access to improved drinking water	8% (1990)	37% (1990)	32% (1990)	33% (1990)	22% (1990)	30% (1990)	66% (1990)	43% (1990)	44% (1990)	39% (1990)
7.9 (U)	Access to Improved sanitation	1% (1990)	4% (1990)	27% (1990)	41% (1990)	23% (1990)	36% (1990)	22% (1990)	22 % (1990)	23% (1990)	40 % (1990)

Table A.8: MVP 2015 Targets, by village

#	Indicator	Koraro, Ethiopia	Bonsasso, Ghana	Sauri, Kenya	Mwandama, Malawi	Tiby, Mali	Pampaida, Nigeria	Mayange, Rwanda	Potou, Senegal	Mbola, Tanzania	Ruhiira, Uganda
1.1	Proportion of population below 1.25 USD (PPP 2005) per day	30.3%	25.6%	19.2%	41.6%	43.1%	31.0%	37.3%	32.9%	36.3%	35.0%
1.2	Poverty Gap ratio	10.6	9.2	7.7	23.0	26.6	15.6	18.5	17.2	14.9	15.2
1.8	Underweight among children under 5 years old	21.8%	11.3%	10.6%	12.8%	17.0%	19.1%	12.4%	13.7%	13.0%	10.0%
2.1	Net attendance ratio in primary education	$\geq 90\%$	$\geq 90\%$	$\geq 90\%$	$\geq 90\%$	$\geq 90\%$	$\geq 90\%$	$\geq 90\%$	$\geq 90\%$	$\geq 90\%$	$\geq 90\%$
2.2	Proportion of pupils starting grade 1 who reach last grade of primary education	$\geq 90\%$	$\geq 90\%$	$\geq 90\%$	$\geq 90\%$	$\geq 90\%$	$\geq 90\%$	$\geq 90\%$	$\geq 90\%$	$\geq 90\%$	$\geq 90\%$
3.1	Gender parity in primary education	$0.97-1.03$	$0.97-1.03$	$0.97-1.03$	$0.97-1.03$	$0.97-1.03$	$0.97-1.03$	$0.97-1.03$	$0.97-1.03$	$0.97-1.03$	$0.97-1.03$
4.1	Under-5 mortality rate (per 1000 births)	64	50	32	81	91	69	54	61	51	53
4.2	Infant mortality rate (per 1000 births)	38	27	22	46	48	32	30	29	32	29
4.3	Measles immunization rate of 1 year-old children	$\geq 90\%$	$\geq 90\%$	$\geq 90\%$	$\geq 90\%$	$\geq 90\%$	$\geq 90\%$	$\geq 90\%$	$\geq 90\%$	$\geq 90\%$	$\geq 90\%$
5.2	Skilled birth attendance	$\geq 70\%$	$\geq 70\%$	$\geq 70\%$	$\geq 70\%$	$\geq 70\%$	$\geq 70\%$	$\geq 70\%$	$\geq 70\%$	$\geq 70\%$	$\geq 70\%$
5.3	Modern contraception use	39.3%	38.5%	54.2%	68.7%	31.1%	28.4%	39.1%	29.2%	50.8%	39.6%
5.5	Antenatal care coverage	$\geq 70\%$	$\geq 70\%$	$\geq 70\%$	$\geq 70\%$	$\geq 70\%$	$\geq 70\%$	$\geq 70\%$	$\geq 70\%$	$\geq 70\%$	$\geq 70\%$
6.7	Children under 5 sleeping under insecticide-bed nets	$\geq 80\%$	$\geq 80\%$	$\geq 80\%$	$\geq 80\%$	$\geq 80\%$	$\geq 80\%$	$\geq 80\%$	$\geq 80\%$	$\geq 80\%$	$\geq 80\%$
7.8	Access to improved drinking water	54%	68.5%	66%	66.5%	61%	65%	83%	71.5%	73%	69.5%
7.9	Access to Improved sanitation	50.5%	52%	63.5%	70.5%	61.5%	68%	61%	61%	61.5%	70%

A.6 Excluded MDG Indicators

Table A.9: Excluded MDG Indicators

#	Indicator
	MDG Goal 1: Eradicate extreme poverty and hunger
1.3	Share of poorest quintile in national consumption
1.4	Growth rate of GDP per person employed
1.5	Employment-to-population ratio
1.6	Proportion of employed people living below $1 (PPP) per day
1.7	Proportion of own-account and contributing family workers in total employment
1.9	Proportion of population below minimum level of dietary energy consumption
	MDG Goal 2: Achieve universal primary education
2.3	Literacy rate of 15-24 year olds, women and men
	MDG Goal 3: Promote gender equality and empower women
3.2	Share of women in wage employment in the non-agricultural sector
3.3	Proportion of seats held by women in national parliament
	MDG Goal 5: Improve maternal health
5.1	Maternal mortality ratio
5.4	Adolescent birth rate
5.5	Antenatal care coverage (at least one visit with skilled health professional)
5.6	Unmet need for family planning
	MDG Goal 6: Combat HIV/AIDS, malaria and other diseases
6.1	HIV prevalence among population aged 15-24 years
6.2	Condom use at last high risk sex
6.3	Proportion of population aged 15-24 years with comprehensive correct knowledge of HIV/AIDS
6.4	Ratio of school attendance of orphans to school attendance of non-orphans aged 10-14 years
6.5	Proportion of population with advanced HIV infection with access to antiretroviral drugs
6.6	Incidence and death rates associated with malaria
6.8	Proportion of children under 5 with fever who are treated with appropriate antimalarial drugs
6.9	Incidence, prevalence, and death rates associated with tuberculosis
6.10	Proportion of tuberculosis cases detected and cured under directly observed treatment short course
	MDG Goal 7: Ensure environmental sustainability
7.1	Proportion of land area covered by forest
7.2	C02 emissions, total, per capita and per $1 GDP (PPP)
7.3	Consumption of ozone depleting substances
7.4	Proportion of fish stocks withing safe biological limits

Continued on next page

Table A.9 – continued from previous page

#	Indicator
7.5	Proportion of total water resources used
7.6	Proportion of terrestrial and marine areas protected
7.7	Proportion of species threatened with extinction
7.10	Proportion of urban population living in slums
	MDG Goal 8: To develop a global partnership for development
8.1	Net ODA, total and to the least developed countries, as percentage of OECD/DAC donors' gross national income
8.2	Proportion of total bilateral, sector-allocable ODA of OECD/DAC donors to basic social services (basic education, primary health care, nutrition, safe water and sanitation)
8.3	Proportion of bilateral official development assistance of OECD/DAC donors that is united
8.4	ODA received in landlocked developing countries as a proportion of their gross national incomes
8.5	ODA received in small island developing States as a proportion of their gross national incomes
8.6	Proportion of total developed country imports (by value and excluding arms) from developing countries and least developed countries, admitted free of duty
8.7	Average tariffs imposed by developed countries on agricultural products and textiles and clothing from developing countries
8.8	Agricultural support estimate for OECD countries as a percentage of their gross domestic product
8.9	Proportion of ODA provided to help build trade capacity
8.10	Total number of countries that have reached their HIPC decision points and number that have reached their HIPC completion points (cumulative)
8.11	Debt relief committed under HIPC and MDRI Initiatives
8.12	Debt service as a percentage of exports of goods and services
8.13	Proportion of population with access to affordable essential drugs on a sustainable basis
8.14	Fixed telephone lines per 100 inhabitants
8.15	Mobile cellular subscriptions per 100 inhabitants
8.16	Internet users per 100 inhabitants

A.7 Impact Evaluation - Technical Details

A.7.1 Small area estimation

There are many types of small area models (Ghosh and Rao, 1994; Ghosh and Natarajan, 1999; Nadram, 2000; Rao, 2003; Jiang and Lahiri, 2006). For continuous variables, we can fit a unit-level linear model,

$$\begin{aligned} y_i &\sim N\left(\mathbf{u}_i^T\boldsymbol{\beta} + \eta_{a[i]}, \sigma_y^2\right) \text{ for individuals } i \\ \eta_a &\sim N\left(\mathbf{x}_a^T\boldsymbol{\gamma} + \kappa_{d[a]}, \sigma_a^2\right) \text{ for EA } a \\ \kappa_d &\sim N\left(\mathbf{z}_d^T\boldsymbol{\delta} + \omega_{r[d]}, \sigma_d^2\right) \text{ for districts } d \\ \omega_r &\sim N\left(0, \sigma_r^2\right) \text{ for regions } r, \end{aligned} \tag{4.1}$$

where $\mathbf{u}_i$ are covariates available in both census and survey, $\mathbf{x}_a$ are EA-level variables, including the size of the EA and population density, $\mathbf{z}_d$ are district-level variables, including size of the district, and information about political institutions and ethnic composition. Geographical variables (e.g. distance to the coast, rainfall, distance to urban areas, distance to a main road) can be included at various levels of the model. We can include more levels between EA and district (e.g. counties or parishes). We need to include all variables used in the sample design, such as the EA size, in order to guarantee ignorability of the data collection mechanism. We may include household effects, or define the households

as the units i. For binary variables, an analogous model to model 4.1 can be fit,

$$\begin{aligned}
y_i &\sim \text{Bern}\,(p_i) \text{ for individuals } i \\
\text{logit}\,(p_i) &= \mathbf{u}_i^T \boldsymbol{\beta} + \eta_{a[i]} \text{ for individuals } i \\
\eta_a &\sim N\left(\mathbf{x}_a^T \boldsymbol{\gamma} + \kappa_{d[a]}, \sigma_a^2\right) \text{ for EA } a \\
\kappa_d &\sim N\left(\mathbf{z}_d^T \boldsymbol{\delta} + \omega_{r[d]}, \sigma_d^2\right) \text{ for districts } d \\
\omega_r &\sim N\left(0, \sigma_r^2\right) \text{ for regions } r.
\end{aligned}$$

Alternatively, instead of unit-level models we can fit a Fay and Herriot (1979) model, where the lowest level of the model is approximated by a non-Bayesian calculation without a complete model for the complex survey data structure (Zaslavsky, 2011):

$$\begin{aligned}
\widehat{\overline{y}}_a &\sim N(\overline{Y}_a, v_a) \text{ for EA } a \\
\overline{Y}_a &\sim N(\mathbf{x}_a^T \boldsymbol{\gamma} + \kappa_{d[a]}, \sigma_a^2) \text{ for EA } a \\
\kappa_d &\sim N(\mathbf{z}_d^T \boldsymbol{\delta} + \omega_{r[d]}, \sigma_d^2) \text{ for districts } d \\
\omega_r &\sim N(0, \sigma_r^2) \text{ for regions } r,
\end{aligned} \tag{4.2}$$

where $\widehat{\overline{y}}_a$ is the standard design-based estimate of the mean in EA a and v_a its sampling variance. Here the v_a account for the sampling design. The $\mathbf{x}_a$ can include EA-level means from the census. For binary variables, an analogous model to model 4.2 can be fit,

$$\begin{aligned}
y_a &\sim \text{Bin}\,(n_a, p_a) \text{ for EA } a \\
\text{logit}\,(p_a) &\sim N(\mathbf{x}_a^T \boldsymbol{\gamma} + \kappa_{d[a]}, \sigma_a^2) \text{ for EA } a \\
\kappa_d &\sim N(\mathbf{z}_d^T \boldsymbol{\delta} + \omega_{r[d]}, \sigma_d^2) \text{ for districts } d \\
\omega_r &\sim N(0, \sigma_r^2) \text{ for regions } r,
\end{aligned}$$

where y_a is the total number of "successes" in area a, and p_a the finite population proportion of successes in area a, which in fact equals $\overline{Y}_a$ for binary y.

Comparing models 4.1 and 4.2, it may be easier to extend 4.2 to jointly model variables (DeSouza, 1992; Datta et al., 1998; Raghunathan et al., 2007; Li and Zaslavsky, 2010), because different variables apply to different individuals, i. We may be more comfortable with the normality assumption in model 4.2, because of the central limit theorem. In either model we may consider transforming the variables to more closely approximate normality (Rao, 2003).

For indicator 3.1 (ratio of girl to boy school attendance), we will model the fraction of school children who are girls, which is a sample mean and more easily modeled than a ratio. An analogous model to model 4.2 can be fit for indicators 2.2 (estimated probability of a student in grade 1 reaching the end of primary school), and the mortality rates 4.1 and 4.2. Instead of $\overline{Y}_a$, we use the estimated probability or mortality rate, call these θ_a. We will make transformations, if necessary, so that the normal approximation is most reasonable.

In model 4.1, if the sampling fraction in area a is small, the finite population mean in area a, $\overline{Y}_a$, is $\overline{\mathbf{U}}_a^T\boldsymbol{\beta} + \eta_a$, where $\overline{\mathbf{U}}_a$ is the population mean of covariates $\mathbf{u}_i$ in EA a, and η_a is estimated as draws of η_a from the posterior if there is survey data in the EA, as $\mathbf{x}_a^T\boldsymbol{\gamma} + \kappa_{d[a]}$ if there is survey data only in the district and not in the EA, and as $\mathbf{x}_a^T\boldsymbol{\gamma} + \mathbf{z}_{d[a]}^T\boldsymbol{\delta} + \omega_{r[d[a]]}$ if there is survey data only in the region and not in the district. Similarly, in model 4.2, if there is survey data in the EA, the finite population mean is estimated with draws of $\overline{Y}_a$ from the posterior, otherwise if there is survey data only in the district, it is estimated with draws of $\mathbf{x}_a^T\boldsymbol{\gamma} + \kappa_{d[a]}$, and if there is only survey data in the region, it is estimated with draws of $\mathbf{x}_a^T\boldsymbol{\gamma} + \mathbf{z}_{d[a]}^T\boldsymbol{\delta} + \omega_{r[d[a]]}$. We can also use the model to get population means at coarser granularities, combining EAs.

We will perform posterior predictive checks and adjust these models appropriately (Gel-

man et al., 2014, Chapters 6 and 7). We may also consider a conditional autoregression (CAR) spatial model to relax our assumptions of exchangeability of EAs within districts and districts within regions (You and Zhou, 2011; Rao, 2003, p.86). We may extend the models by allowing the slope parameters to vary by area or larger region.

Complications

Our problem is more complicated than described above, for several reasons.

For privacy reasons, the DHS reports displaced latitude and longitude of the EAs it samples, by up to 5 km in rural areas (Measure DHS/ICF International, 2012; DHS, 2014). We should be able to identify that the sampled EA is one of a few EAs from the census, since the average size of an EA is 5-20 km^2 (from a crude computation of country area divided by total number of EAs). We can modify the models above by defining a small area to be a group of a few nearby EAs. If the census data shows a smoothness across nearby EAs, with neighboring EAs having similar values for the census variables, the DHS displacement is not problematic. Alternatively, we may be able to get access to some non-displaced DHS data, depending on privacy restrictions.

We may not be able to get access to census data that is georeferenced at the EA-level, but rather, only at larger sub-district administrative area (called by different names in different countries). In this case, the $\mathbf{x}_a$ variables would be constant for all EAs within the same administrative area.

Another complication is that in general, censuses were not done at the same times as the DHS, and neither at the same time as the project start dates, see Table A.10.

For example, in Kenya, censuses were done in 1999 and 2009, and DHS in 2003 and 2008. In the unit-level small area model 4.1, this is problematic because the DHS covariates $\mathbf{u}_i$

Table A.10: Timing of the DHS and country censuses.

MV	Start date	DHS dates	census dates
Koraro, Ethiopia (EK)	Q1 2005	2000, 2005, 2011	1994, 2007
Bonsasso, Ghana (GB)	Q3 2006	2003, 2008	2000, 2010
Sauri, Kenya (KS)	Q1 2005	2003, 2008-9	1999, 2009
Mwandama, Malawi (MM)	Q3 2006	2000, 2004, 2010	1998, 2008
Tiby, Mali (MT)	Q3 2006	2001, 2006	1998, 2009
Pampaida, Nigeria (NP)	Q2 2006	2003, 2008	1991, 2006
Mayange, Rwanda (RM)	Q3 2006	2000, 2005, 2010	2002, 2012
Potou, Senegal (SP)	Q2 2006	2005, 2010-11	2002, 2013
Mbola, Tanzania (TM)	Q2 2006	2004-5, 2010	2002, 2012
Ruhiira, Uganda (UR)	Q2 2006	2000-1, 2006, 2011	2002, 2013

are from a different time than the census averages $\overline{\mathbf{U}}_a$. In the area-level model 4.2, this is not a problem, but the further in time the census is from the DHS, the less predictive census variables are likely to be. One option is to linearly interpolate census data between 1999 and 2009 to get 2003 or 2008 values for the covariates (this interpolation can be assessed using DHS data in 2003 and 2008). Alternatively, we could use the 1999 or 2009 data as the covariates. We propose to do both, and compare results.

Use of small area estimates in the selection of comparison areas

In the selection of comparison areas, we have two options. The first is to match comparison areas to MVs using only estimates from the small area models, based on DHS and census data. The second is to match estimates from the small area models to MV project data collected at baseline.

The first approach avoids a possible lack of comparability between DHS data and project data. For example, in Kenya, we would match on 2003 estimates from the small area models. We can also match on 1999 census data alone, avoiding the noise from small area estimation. The disadvantage to using only 1999 census data is that several key variables

are not recorded by censuses, and changes between 1999 and 2005 may turn good-looking matches in 1999 into poor matches in 2005. Matching on small areas estimates in 2003, we will have noisy estimates for both the MVs and comparison areas.

We can avoid noisy estimates in the MVs by relying on project data, using the second approach. In this approach, in Kenya we would linearly interpolate 2003 and 2008 small area estimates to get 2005 estimates in comparison areas, and match these to MV project data at baseline. This method is less robust than the first because it relies both on the linear interpolation and on the comparability of DHS to project data. We can assess the comparability of DHS to project data most easily in countries for which the DHS was done near project baseline. We can assess the linear interpolation on variables collected by both the census and DHS by fitting a curve to the time points from a few DHS and census rounds.

Another possible objection to the second approach is the use of post-treatment data. However, we do not condition on post-treatment data, but rather, we condition on a *function* of it, as recommended by Liu and Meng (2014). In fitting small area models with post-treatment data, we leave out the MV and nearby areas, because those areas are not exchangeable with other areas in the country.

In addition to matching treatment and control areas on pre-treatment levels, we would ideally also match on trends. We want to match MVs on an upswing with areas also on an upswing. We can get an approximate trend in comparison areas from the slopes from the linear interpolations discussed above. For the MVs, we can compute the slope between the MV baseline and an SAE estimate from a pre-baseline DHS (for example, in 2003), assuming we believe in the comparability of DHS and project data. Otherwise, if there are multiple pre-treatment DHS, we can compute slopes from those.

A.7.2 Propensity Score Model

We require estimates of propensity scores. We follow the conventional approach in the literature and use logistic regression on our variables unaffected by MVP (Imbens and Rubin, 2014). The procedures in A.7.1 give small-area-level baselines (and possibly also trends) of indicators measured by the DHS. There will be substantial uncertainty in these baselines, which we will address in A.7.3. For the MVs, we also have baselines, which may come from small area estimates or from project data (see A.7.1). We will fit a propensity score model,

$$\begin{aligned} \text{logit}(P(Z_a = 1)) &= \mathbf{x}_a\boldsymbol{\delta} + \kappa_{d[a]} \text{ for EA } a \\ \kappa_d &\sim N(\mathbf{z}_d^T\boldsymbol{\delta} + \omega_{r[d]}, \sigma_d^2) \text{ for districts } d \\ \omega_r &\sim N(0, \sigma_r^2) \text{ for regions } r, \end{aligned} \tag{4.3}$$

where Z_a is a treatment indicator for area a, $\mathbf{x}_a$ is a vector including the baselines from the small area models in A.7.1 and geographical variables, and the $\mathbf{z}_d$ are district-level variables. Covariates can include those mentioned above, in A.7.1. Model specification will be guided by an iterative process of improving the model using posterior predictive checks (Gelman et al., 2014, Chapters 6 and 7).

A.7.3 Candidate Models for Causal Inference

Here we suggest a few types of causal models that we propose to fit to the end-line outcome data. Let j index a village, either an MV or a comparison village. These are nested into ten countries, indexed by c. Let Z_j be the indicator of treatment for village j. Let $\theta_j^{(k,t)}$ denote the village-level indicator for outcome k at time t. For the binary and continuous

outcomes, this is the finite population mean, $\overline{Y}_j^{(k,t)}$, of individual-level outcomes $y_i^{(k,t)}$. For indicator 2.2, $\theta_j^{(k,t)}$ is the probability of completing primary school, for indicator 3.1 it is the ratio of girl to boy school attendance, and for indicators 4.1 and 4.2, the under-5 and infant mortality rates.

Let $\mathbf{x}_j$ be a vector of village-level covariates, including estimated baseline variables from the models in A.7.1. Some villages may span more than one small area, so we will aggregate appropriately. The causal models are fit conditional on these baselines, whose substantial uncertainty we discuss in A.7.3.

Our models are similar to the multilevel models for matched-pair cluster designs suggested by Hill and Scott (2009). Where not otherwise specified, priors on parameters are non-informative.

Our estimands are superpopulation average treatment effects, conditional on covariates (Gelman et al., 2014, Chapter 8). Thus, we imagine that the villages were "sampled" from a population of villages with similar covariates (similar extreme poverty in 2005), with high levels of political buy-in, where MVP treatment would not have been disrupted by financing shortages or political instability (See Section 4.2.1 in the main body of the paper).

Single-outcome models

We consider a ladder of models, starting with simple models, and building to more complex models. The first few rungs of the ladder include only one outcome at a time, and treatment effects that do not vary across countries. The first rung of the ladder includes no covariates, and totally pools across countries. For each outcome k that is continuous (income, weight-for-age-z-score, and number of antenatal care visits) we will fit a linear

model,

$$y_i^{(k,2015)} \sim N\left(\mu^{(k)} + \tau^{(k)} Z_{j[i]}, \sigma_y^2\right) \text{ for individuals } i.$$

The second rung of the ladder includes no covariates but does include partially-pooling over the countries:

$$\begin{aligned} y_i^{(k,2015)} &\sim N\left(\mu^{(k)} + \tau^{(k)} Z_{j[i]} + \alpha_{c[j[i]]}^{(k)}, \sigma_y^2\right) \text{ for individuals } i \\ \alpha_c^{(k)} &\sim N(0, \sigma_\alpha^2) \text{ for countries } c = 1, ..., 10. \end{aligned}$$

The third rung of the ladder includes covariates as well:

$$\begin{aligned} y_i^{(k,2015)} &\sim N\left(\mu^{(k)} + \tau^{(k)} Z_{j[i]} + \mathbf{x}_{j[i]}\boldsymbol{\delta}^{(k)} + \alpha_{c[j[i]]}^{(k)}, \sigma_y^2\right) \text{ for individuals } i \\ \alpha_c^{(k)} &\sim N(0, \sigma_\alpha^2) \text{ for countries } c = 1, ..., 10. \end{aligned} \tag{4.4}$$

For binary outcomes, we will fit analogous logistic models,

$$\begin{aligned} \text{logit}\left(P(y_i^{(k,2015)} = 1)\right) &= \mu^{(k)} + \tau^{(k)} Z_{j[i]} + \mathbf{x}_{j[i]}\boldsymbol{\delta}^{(k)} + \alpha_{c[j[i]]}^{(k)} \text{ for individuals } i \\ \alpha_c^{(k)} &\sim N(0, \sigma_\alpha^2) \text{ for countries } c = 1, ..., 10. \end{aligned} \tag{4.5}$$

For the mortality outcomes (indicators 4.1 and 4.2), see the next section.

We can also fit village-level models. Let $\widehat{\theta}_j^{(k,2015)}$ denote the estimated village-level indicator, and v_j its variance. Then we can fit a model:

$$\begin{aligned} \widehat{\theta}_j^{(k,2015)} &\sim N\left(\mu^{(k)} + \tau^{(k)} Z_j + \mathbf{x}_j\boldsymbol{\delta}^{(k)} + \alpha_{c[j]}^{(k)}, v_j\right) \text{ for villages } j \\ \alpha_c^{(k)} &\sim N(0, \sigma_\alpha^2) \text{ for countries } c = 1, ..., 10. \end{aligned} \tag{4.6}$$

We may need to transform some variables to make the normality assumption more plausible (e.g. logit transform the proportions). Without individual-level covariates, $\tau^{(k)}$ has the same interpretation in either model 4.4 or 4.6. However, model 4.4 makes the assumption of normality of the individual outcomes rather than the aggregates.

We will also consider interactions between treatment and covariates to assess sensitivity to the assumption that the slopes ($\boldsymbol{\delta}^{(k)}$) vary by treatment group. However, we may not have the precision to be able to estimate these interactions without some strong regularization via prior distributions.

Mortality outcomes - Survival models

For indicators 4.1 and 4.2, standard methods used by the DHS are described in Rutstein and Rojas (2006, p.99-101). We can use these standard methods to compute village-level mortality rates, and fit model 4.6. Alternatively, we can fit a survival model. For indicator 4.1 (under-5 mortality rate), the relevant end-line study period is 2010-2015, following the conventions in Rutstein and Rojas (2006); MDG (2014). Through women's birth histories collected in 2015, we will have birth and death dates (if the child died) for any children age 0-5 years alive during this study period. The complications with considering under-5 mortality in 2010-2015 are: we want a child born before 2010 to contribute to the analysis only during the study period, and we want only ages 0-5 to contribute to the analysis. To accomplish this we propose the following method:

Let J_{0i} be child i's *joining time*, which equals 2010 for children born before 2010, and equals the calendar year of birth for children born after 2010. Let A_{0i} be child i's *age adjustment*, which equals the child's age in 2010 for children born before 2010, and equals zero for children born after 2010. Let T_i^* be child i's survival time, i.e. how many years child i lives in total. Then $T_i = T_i^* - A_{0i}$ is the survival time since the joining time J_{0i}. The censoring

time is $C_i = \min(5 - A_{0i}, 2015 - J_{0i})$ because children born before 2010 are censored when they reach age 5 and those born after 2010 are censored in 2015. The observed data are $(U_i, \delta_i, \mathbf{X}_i)$ where $U_i = \min(T_i, C_i)$, $\delta_i = I(T_i \leq C_i)$, and $\mathbf{X}_i$ are covariates, including treatment indicator, country effect, and other variables.

Now, C_i may be dependent on T_i because both may depend on A_{0i}: for people born before 2010, $C_i = 5 - A_{0i}$ while $T_i = T_i^* - A_{0i}$, and T_i can also be affected by A_{0i} through interaction with treatment because of prior beneficial effect of treatment leading to improved survival during the study period. Thus, we want to condition on A_{0i} in our analysis so that C_i and T_i are more likely to be independent. We also want to condition on J_{0i} because otherwise $(U_i, \delta_i, \mathbf{X}_i)$ may not be i.i.d. (independent and identically distributed): for a child with a smaller value of J_{0i}, the observation $(U_i, \delta_i, \mathbf{X}_i)$ is more likely to be $(T_i, 1, \mathbf{X}_i)$, while for a child with a larger value of J_{0i} (but same value of covariates $\mathbf{X}_i$), the observation $(C_i, 0, \mathbf{X}_i)$ is more likely. In addition to including A_{0i} and J_{0i} as covariates, we need to include the interaction of A_{0i} and treatment ($Z_{j[i]}$) in order to account for the possible benefits that children born before 2010 may have had from getting the treatment for a few years prior to joining the study period.

Finally, we fit a survival analysis model (Cox, 1972; Ibrahim et al., 2001) adjusting for the variables mentioned above, with country effects as in the models in A.7.3. The coefficient of treatment, $\tau^{(k)}$, represents a log hazard ratio, comparing the hazard of death among children in a treatment village to those in a comparison village, among children with the same covariates adjusted for in the model, ages 0-5 during 2010-2015. We can also use the model to compute other summaries of the treatment effect (including the difference or ratio of the probability of a child surviving to age 5 in treatment versus comparison villages) by estimating the baseline survivor function.

Joint-outcome models

If we wish to model all outcomes in a joint model, care is needed because the outcomes apply to different populations (e.g. children, infants, or women). One option is to fit village-level outcomes,

$$\begin{bmatrix} \widehat{\theta}_j^{(1,2015)} \\ \vdots \\ \widehat{\theta}_j^{(k,2015)} \\ \vdots \\ \widehat{\theta}_j^{(14,2015)} \end{bmatrix} \sim N \left(\begin{bmatrix} \mu^{(1)} + \tau^{(1)} Z_j + \mathbf{x}_j \boldsymbol{\delta}^{(1)} + \alpha_{c[j]}^{(1)} \\ \vdots \\ \mu^{(k)} + \tau^{(k)} Z_j + \mathbf{x}_j \boldsymbol{\delta}^{(k)} + \alpha_{c[j]}^{(k)} \\ \vdots \\ \mu^{(14)} + \tau^{(14)} Z_j + \mathbf{x}_j \boldsymbol{\delta}^{(14)} + \alpha_{c[j]}^{(14)} \end{bmatrix}, \Sigma_j \right) \text{ for villages } j$$

$$\begin{bmatrix} \alpha_c^{(1)} \\ \vdots \\ \alpha_c^{(k)} \\ \vdots \\ \alpha_c^{(14)} \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ \vdots \\ 0 \\ \vdots \end{bmatrix}, \Sigma_{country} \right) \text{ for countries } c = 1, ..., 10. \tag{4.7}$$

We may add a level to assign a multivariate normal prior to $\left(\mu^{(k)}, \tau^{(k)}, \boldsymbol{\delta}^{(k)}\right)$, with some grouping of outcomes k to relax exchangeability assumptions. Both covariance matrices Σ_j and $\Sigma_{country}$ will have block structures to reflect the different MDGs. Entries in matrix Σ_j contain the sampling variances of each $\widehat{\theta}_j^{(k,2015)}$, which account for household clustering.

We will perform posterior predictive checks on this model to assess its fit to the data (Gelman et al., 2014, Chapters 6 and 7). We may need to transform variables, and consider latent variables or copulas.

We may also be able to fit an analogous individual-level model, with care given to the fact

that for each individual i, a different set of outcomes is defined. For example, if individual i is a 45-year-old man, he does not have a bednet outcome.

Alternatively, we can also compute the average effect size (AES) estimates across outcomes, following O'Brien (1984); Clingingsmith et al. (2009).

Difference-in-Differences methods

Previous evaluations of the MVP, Clemens and Demombynes (2011) and Pronyk et al. (2012), as well at the proposal for the new northern Ghana MV evaluation, ITAD (2013), use the impact evaluation method of *difference-in-differences*. Difference-in-differences uses measurements at two time points, baseline and end-line (and possibly also time points in between), and an assumption of *additivity* to difference out time-invariant village effects and identify the effect of treatment. Additivity requires the potential gains over time to be the same across treatment and control groups, adjusted for covariates.

Instead of additivity, our models above, often known as *ANCOVA models*, assume unconfoundedness given the baseline outcome variables (at an aggregate village level), and other covariates. Difference-in-differences and ANCOVA models each make different assumptions, neither makes strictly fewer assumptions than the other (Imbens and Wooldridge, 2009, p.70). However, Imbens and Wooldridge (2009) do suggest that the unconfoundedness given baseline approach is, in general, more attractive. In order to test the sensitivity to these assumptions, we propose to fit difference-in-differences models analogous to our above models, for any outcome k for which we have a baseline. If there are large discrepancies between the two types of models, we will have to conclude that we are uncertain which to trust.

Without enough reliable individual-level data at baseline, we cannot fit an individual-

level difference-in-differences model. Thus, we will use village-level baselines, $\theta_j^{(k,2005)}$, whose uncertainty we discuss in A.7.3. For example, an analogous model to model 4.6 is

$$\widehat{\theta}_j^{(k,2015)} - \theta_j^{(k,2005)} \sim N(\mu^{(k)} + \tau^{(k)} Z_j + \mathbf{x}_j^{(-k)} \boldsymbol{\delta}^{(k)} + \alpha_{c[j]}^{(k)}, v_j) \text{ for villages } j$$
$$\alpha_c^{(k)} \sim N(0, \sigma_\alpha^2) \text{ for countries } c = 1, ..., 10,$$

where $\mathbf{x}_j^{(-k)}$ is a vector including baseline outcomes for all k' not equal to k.

In contrast, ITAD (2013) plans to gather panel data on individuals from the MV and the surrounding comparison areas. Panel data on individuals allows ITAD (2013) to difference out time-invariant individual effects (we can only difference out time-invariant village effects), and to adjust for individual-level covariates, which increase both efficiency and validity compared to our approach that only uses village-level information. Without individual-level panel data in comparison areas, we cannot adopt their approach.

Varying Treatment Effects

We plan to fit our above models allowing for treatment effects to vary by Millennium Village, with partial pooling (Hill and Scott, 2009; Feller and Gelman, 2014). For example, extending our model 4.4,

$$y_i^{(k,2015)} \sim N\left(\mu^{(k)} + \tau_c^{(k)} Z_{j[i]} + \mathbf{x}_{j[i]} \boldsymbol{\delta}^{(k)} + \alpha_{c[j[i]]}^{(k)}, \sigma_y^2\right) \text{ for individuals } i \tag{4.8}$$
$$\begin{bmatrix} \alpha_c^{(k)} \\ \tau_c^{(k)} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ \tau \end{bmatrix}, \begin{bmatrix} \sigma_\alpha^2 & \sigma_{\alpha\tau} \\ \sigma_{\alpha\tau} & \sigma_\tau^2 \end{bmatrix}\right) \text{ for countries } c = 1, ..., 10.$$

If we only have one matched comparison village per treatment village, each village's treatment effect cannot be interpreted causally, because we cannot separately identify within-pair variation from treatment effect variation. However, we may want to test for

(and report) this extra heterogeneity, whatever its true source. In A.7.3 we discuss what can be done with more than one comparison village matched to a treatment village.

Multiple Comparisons and Fishing

We have fourteen primary outcomes of interest, discussed in Section 4.5 in the main body of the paper. Inevitably, with multiple comparisons, there will be some that reach the "statistical significance" threshold and some that do not. We will report all intervals of uncertainty and not focus on statistical significance as a summary. To avoid fishing, we will report and compare all results, not selecting only those that are significant and favorable (Humphreys et al., 2013).

Gelman et al. (2012) recommend jointly modeling outcomes of interest, as we discussed in A.7.3. This way, the treatment effects have a 95% probability of collectively being in their 95% intervals obtained by the model (we will be working within the Bayesian paradigm, where parameters have probability distributions). This helps to control the overall type I error rate, if this is a concern.

Accounting for uncertainty in village-level baselines

There will be substantial uncertainty in the village-level baselines, and we wish to propagate this uncertainty in our analysis so that our intervals for the treatment effects honestly reflect the uncertainty in our procedure. We propose to account for uncertainty in baseline value $\theta_j^{(k,2005)}$, by adding a level to the hierarchical propensity score and causal models: $\theta_j^{(k,2005)} \sim N\left(\widehat{\theta}_j^{(k,2005)}, v_j^{(k,2005)}\right)$, where $v_j^{(k,2005)}$ is the posterior variance from the small area estimation procedures described in A.7.1 (see Gelman et al. (2014, p.474) for a similar example). If we choose to use project baseline data (see A.7.1), $v_j^{(k,2005)}$ is the

posterior variance from a simple model that reflects the sampling done within each MV at baseline. We may transform $\theta_j^{(k,2005)}$ to make normality more plausible. For example, for binary indicators we may use a logit transformation.

For areas with a lot of baseline uncertainty, their propensity scores should exhibit a great deal of variability, making them a poor choice for a comparison area.

Assessing Unconfoundedness

Although unconfoundedness cannot be directly tested, there are analyses that can assess its plausibility (Altonji et al., 2005; Imbens and Rubin, 2014, Chap.12). As these tests are done before outcome data are available, we propose to include them in our publicly released design analysis discussed in Section 4.12 in the main body of the paper, if baseline data in comparison areas are sufficiently rich. Methods of assessment include estimating the effect of treatment on a pretreatment outcome, and estimating the causal effect of a treatment known not to have an effect, using multiple control groups.

In our final evaluation report, we propose to do a sensitivity analysis in the style of Rosenbaum and Rubin (1983a); Imbens (2003); Rosenbaum (2005).

Multiple controls per treatment village

If more than one decent comparison village can be identified within one district, and the funding exists, we propose to collect comparison data from a few comparison villages per treatment village. We propose to slightly modify our above regressions, because with multiple comparison villages per treatment village, we can now estimate the variability

in village effects. For example, instead of model 4.8, we would fit the following model:

$$
\begin{aligned}
y_i^{(k,2015)} &\sim N\left(v_{j[i]}^{(k)}, \sigma_y^2\right) \text{ for individuals } i \\
v_j^{(k)} &\sim N\left(\mu^{(k)} + \tau_c^{(k)} Z_j + \mathbf{x}_j \boldsymbol{\delta}^{(k)} + \alpha_{c[j]}^{(k)}, \sigma_v^2\right) \text{ for villages } j \\
\begin{bmatrix} \alpha_c^{(k)} \\ \tau_c^{(k)} \end{bmatrix} &\sim N\left(\begin{bmatrix} 0 \\ \tau \end{bmatrix}, \begin{bmatrix} \sigma_\alpha^2 & \sigma_{\alpha\tau} \\ \sigma_{\alpha\tau} & \sigma_\tau^2 \end{bmatrix}\right) \text{ for countries } c = 1, ..., 10.
\end{aligned}
$$

If σ_v is very small, might we feel comfortable interpreting τ_c as the treatment effect for the Millennium Village in country c, as opposed to in a matched pair design, as discussed in A.7.3.

In addition, we propose to explore, for each village, the creation of a *synthetic control*, a weighted average of comparison villages that serves as a better control than each village alone (Abadie et al., 2010). Synthetic control methods in Abadie et al. (2010) rely on multiple pretreatment measurements to (under some conditions) match unobserved village effects, enabling estimation of the treatment effect for each MV, separately. If enough georeferenced DHS and country censuses can be obtained for multiple pretreatment times, we propose to explore synthetic control methods. We will make a decision about whether these methods are appropriate before we view any outcomes, and release our recommendations in the report discussed in Section 4.12 in the main body of the paper.

Without multiple pretreatment time periods, but with multiple matched treatment and control villages, we should be able to estimate an average treatment effect across all the MVs, since the differences in unobserved village effects between treated and control villages are assumed to have mean zero under unconfoundedness.

A.7.4 Power Calculations

We perform power calculations using simulation-based methods recommended in Gelman and Hill (2007). We do power calculations for four variables: annualized consumption (a measure of income), weight for age z-score, measles immunization, and bednet usage.

We use data from years 0 and 5 (i.e. 2005 and 2010) in the MVs in order to create realistic simulated data as follows. First, we fit modifications of models in A.7.3. For continuous outcomes (annualized consumption, weight for age z-score), we fit a linear model,

$$\begin{aligned} y_i^{(k,2010)} &\sim N\left(\mu^{(k)} + \mathbf{x}_{j[i]}\boldsymbol{\delta}^{(k)} + \alpha_{j[i]}^{(k)}, \sigma_y^2\right) \text{ for individuals } i \\ \alpha_j^{(k)} &\sim N(0, \sigma_\alpha^2) \text{ for MVs } j = 1, ..., 10. \end{aligned} \tag{4.9}$$

For the binary outcomes (measles immunization, bednet usage), we fit a logistic model,

$$\begin{aligned} \text{logit}\left(P(y_i^{(k,2010)} = 1)\right) &= \mu^{(k)} + \mathbf{x}_{j[i]}\boldsymbol{\delta}^{(k)} + \alpha_{j[i]}^{(k)} \text{ for individuals } i \\ \alpha_j^{(k)} &\sim N(0, \sigma_\alpha^2) \text{ for MVs } j = 1, ..., 10. \end{aligned} \tag{4.10}$$

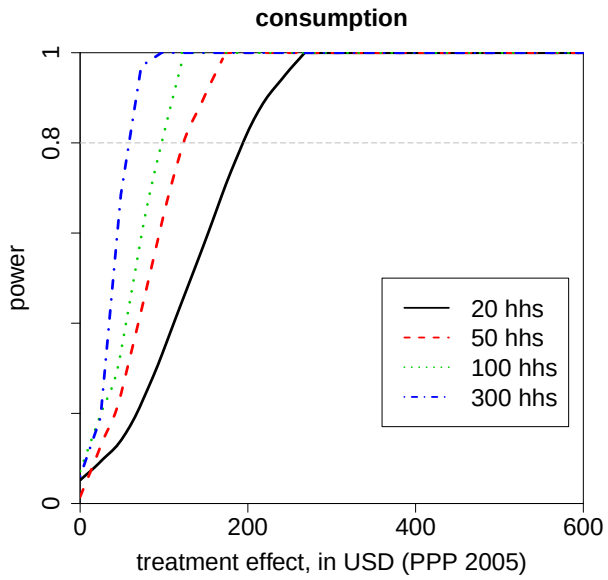
We do power calculations assuming only one matched comparison village per MV, to be conservative. For each pair, one MV and one comparison village, we simulate baselines values from a normal distribution, centered at the real MV data baselines, with standard deviation equal to 10% of the MV baseline, in order to simulate imperfect matching. In the power calculations here we do not account for the uncertainty in these baseline measurements. We take σ_α to be the 50% posterior quantile from model 4.9 or 4.10, and σ_y to be the 50% posterior quantile from model 4.9. (Results for when σ_α is set as the 80% posterior quantile were very similar to those included in this report.)

Using these values for the variance parameters, the simulated baselines, and the posterior 50% quantiles of parameters $\mu^{(k)}, \boldsymbol{\delta}^{(k)}$ from fitting either model 4.9 or 4.10, we generate data for continuous outcomes from model 4.4, and for binary outcomes from model 4.5, taking $\tau^{(k)}$ to be a variety of plausible values for each outcome k. We compute, via simulation, the power (the probability that the estimated treatment effect is statistically significant) for each value of $\tau^{(k)}$ and either 20, 50, 100, or 300 individuals per village. To avoid adjusting for household clustering, we make the conservative assumption that each household provides an effective sample size of one individual, so our samples can be regarded as households (hhs) sampled per village. We run 200 simulations per set of parameter values and sample size.

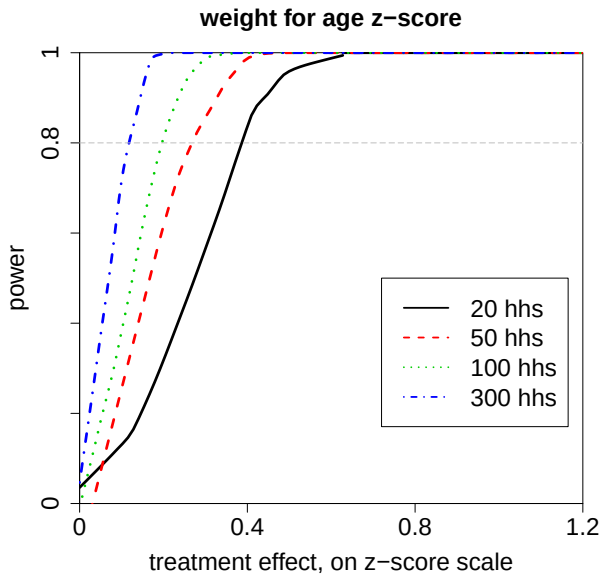
For continuous outcomes, we look at treatment effects ranging from zero to the standard deviation of the outcomes in year five. For binary outcomes, we look at log odds ratios ranging from zero to 1.5. We fit models 4.4 and 4.5 to the data (generated from these same models) to obtain estimates of treatment effects in each simulation.

In Figure A.19 we plot power as a function of treatment effect for four outcomes: annualized consumption, weight for age z-score, measles immunization, and bednet usage. In Figure A.20, we plot the *Type S* error, the probability that the estimated treatment effect has the incorrect *sign*, if it is statistically significant. In Figure A.21, we plot the *Type M* error, the expected absolute value of the estimate divided by the true effect size, if it is statistically significant. Type M error captures the error *magnitude* (Gelman and Carlin, 2013).

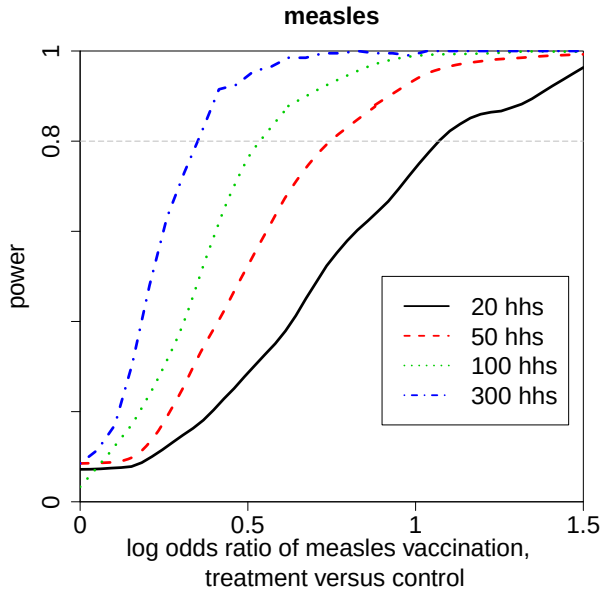
Results for the difference-in-differences version of models 4.4 and 4.5 yielded similar results. The usual gains in efficiency from ANCOVA models (as compared to difference-in-differences, see McKenzie (2012)) were not seen here, likely because the baselines are not at the individual level, but rather, at the village level.



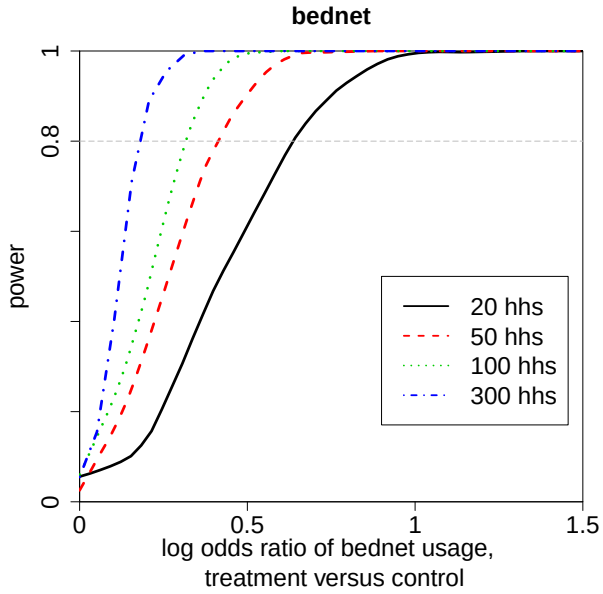
(a) Power for annualized consumption.



(b) Power for weight for age z-score.

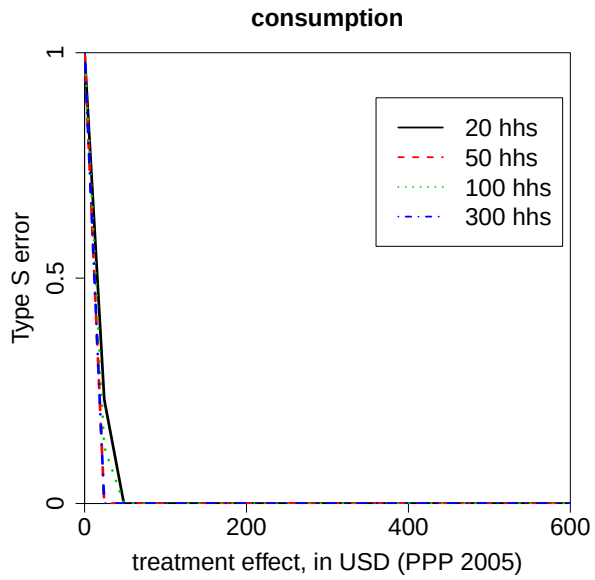


(c) Power for measles immunization.

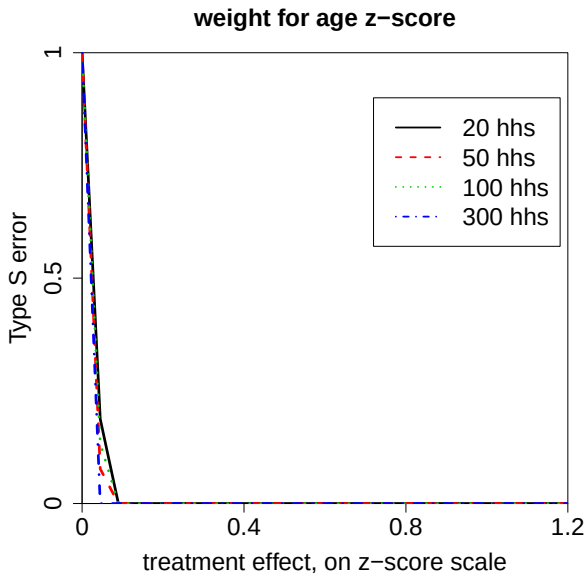


(d) Power for bednet usage.

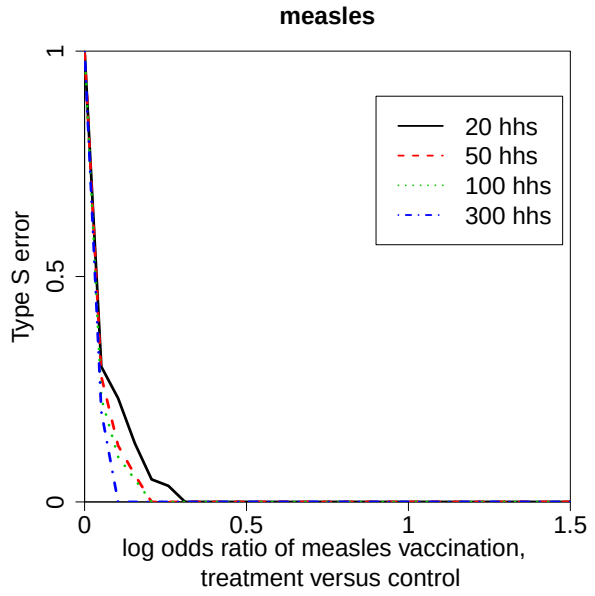
Figure A.19: Power (the probability that the estimated treatment effect is statistically significant) as a function of treatment effect for four different outcomes: (a) annualized consumption, in USD (PPP 2005), (b) weight for age z-score, (c) measles immunization, (d) bednet usage; and four different sample sizes: 20, 50, 100, 300 households (hhs). We fit a model that assumes unconfoundedness given baseline outcomes.



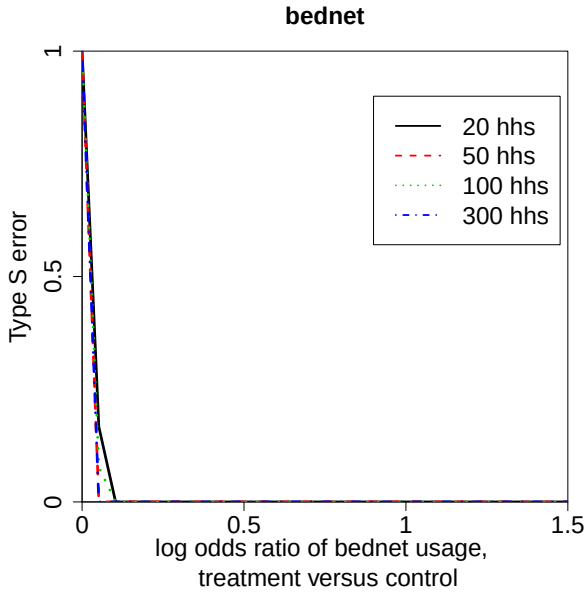
(a) Type S error probabilities for annualized consumption.



(b) Type S error probabilities for weight for age z-score.

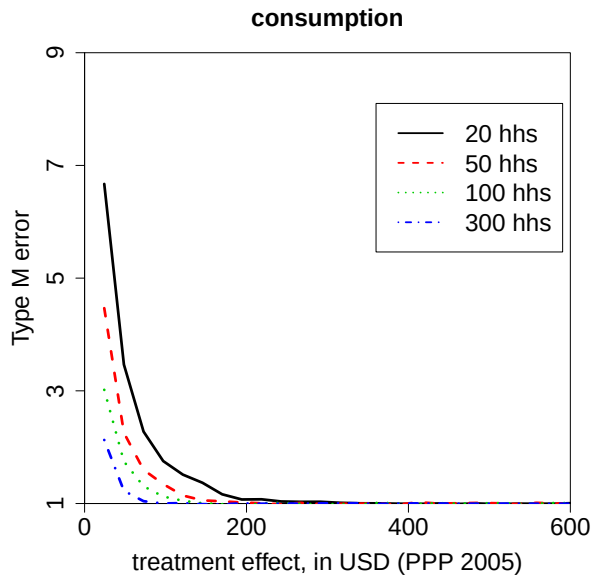


(c) Type S error probabilities for measles immunization.

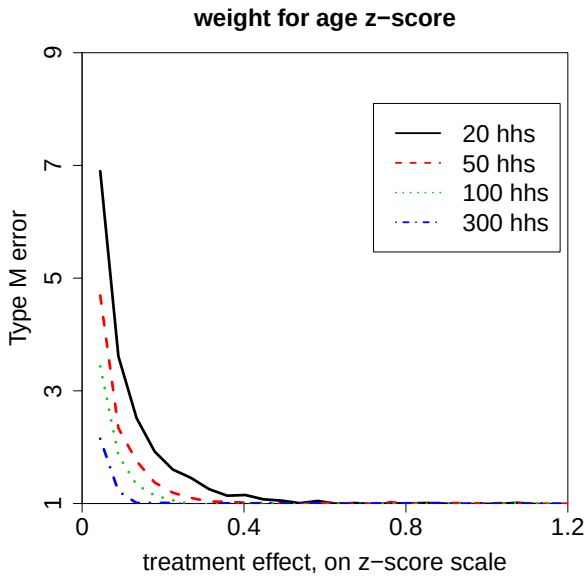


(d) Type S error probabilities for bednet usage.

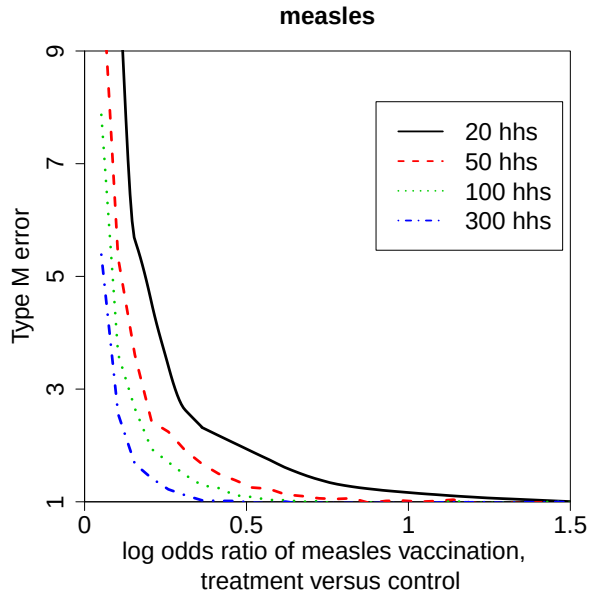
Figure A.20: Type S error probabilities (the probability that the estimated treatment effect has the incorrect sign, if it is statistically significant) as a function of treatment effect for four different outcomes: (a) annualized consumption, in USD (PPP 2005), (b) weight for age z-score, (c) measles immunization, (d) bednet usage; and four different sample sizes: 20, 50, 100, 300 households (hhs). We fit a model that assumes unconfoundedness given baseline outcomes.



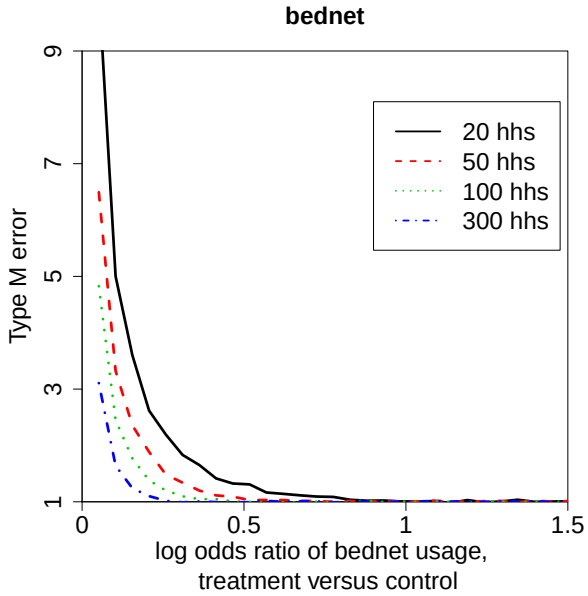
(a) Type M error for annualized consumption.



(b) Type M error for weight for age z-score.



(c) Type M error for measles immunization.



(d) Type M error for bednet usage.

Figure A.21: Type M errors (the expected absolute value of the estimate divided by the true effect size, if it is statistically significant) as a function of treatment effect for four different outcomes: (a) annualized consumption, in USD (PPP 2005), (b) weight for age z-score, (c) measles immunization, (d) bednet usage; and four different sample sizes: 20, 50, 100, 300 households (hhs). We fit a model that assumes unconfoundedness given baseline outcomes.

Simplifications

In this power analysis we simplify the real design in the following ways:

1. We assume fairly good matching on baseline variables.
2. We assume perfect measurement of baseline variables.
3. We assume data are generated from the same model that we fit, which assumes constant treatment effect across villages.
4. We assume a simple matched pair design with one comparison village matched to each research village (MV1). This does not take into account any use of multiple comparison villages per treatment village, as discussed in A.7.3.
5. We assume no nonresponse.

A.7.5 Acknowledgements

Avi Feller, Jennifer Hill, Abhishek Chakrabortty, Keli Liu, Peng Ding, Natalie Bau, Natalie Exner, Erin Doxsey-Whitfield, Susana B. Adamo, Marcia Castro, Nathalie Mumaw, Sehrish Bari, Kytt MacManus, Kyle DeRosa, Ryan Marriott, Teddy Svoronos, Matthew Harris, Christopher Blattman, Dean Karlan, Macartan Humphreys.

References

(2010). "Harvests of development: the millennium villages after three years," Technical report, The Earth Institute at Columbia University, New York, NY.

(2014). URL `http://mdgs.un.org/unsd/mdg/Metadata.aspx`.

(2014). URL `http://www.measuredhs.com/faq.cfm`.

Abadie, A., A. Diamond, and J. Hainmueller (2010). "Synthetic control methods for comparative case studies: Estimating the effect of california's tobacco control program," *Journal of the American Statistical Association*, 106, 493–505.

Abadie, A. and G. W. Imbens (2011). "Bias-corrected matching estimators for average treatment effects," *Journal of Business and Economic Statistics*, 29, 1–11.

Agresti, A. (1994). "Simple capture-recapture models permitting unequal catchability and variable sampling effort," *Biometrics*, 50, 494–500.

Agresti, A. (2002). *Categorical Data Analysis*, John Wiley and Sons, 2 edition.

Ahren, P. (1976). *Economic evaluation methods in community planning*, Swedish Council for Building Research.

Altonji, J. G., T. E. Elder, and C. R. Taber (2005). "Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools," *Journal of Political Economy*, 113, 151–184.

Angrist, J. D. and J. S. Pischke (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton University Press.

Balk, D., T. Pullum, A. Storeygard, R. Greenwell, and M. Neuman (2004). "A spatial analysis of childhood mortality in West Africa," *Population, Space and Place*, 10, 175–216.

Balk, D., A. Storeygard, M. Levy, J. Gaskell, M. Sharma, and R. Flor (2005). "Child hunger in the developing world: An analysis of environmental and social correlates," *Food Policy*, 30, 584–611.

Bang, H. and J. M. Robins (2005). "Doubly robust estimation in missing data and causal inference models," *Biometrics*, 61, 962–972.

Baron, R. M. and D. A. Kenny (1986). "The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations," *Journal of Personality and Social Psychology*, 51, 1173–82.

Bartolucci, F. and A. Forcina (2001). "Analysis of capture-recapture data with a rasch-type model allowing for conditional dependence and multidimensionality," *Biometrics*, 57, 714–719.

Bartolucci, F. and A. Forcina (2006). "A class of latent marginal models for capture-recapture data with continuous covariates," *Journal of the American Statistical Association*, 101, 786–794.

Bergsma, P. and T. Rudas (2002). "Marginal models for categorical data," *Annals of Statistics*, 30, 140–159.

Bicego, G. and O. B. Ahmad (1996). "Demographic and health surveys: Infant and child mortality," Technical report, Demographic and Health Surveys, Calverton, Maryland, URL `https://dhsprogram.com/pubs/pdf/CS20/00FrontMatter00.pdf`.

Bishop, Y. M. M., S. Fienberg, and P. H. Holland (1975). *Discrete Multivariate Analysis: Theory and Practice*, Cambridge, MA: MIT Press.

Blattman, C. (2007). "Do the millenium villages work?" Blog post, URL `http://chrisblattman.com/2007/12/28/do-the-millenium-villages-work/`.

Blattman, C. (2009). "Am i actually sticking up for the millennium villages?" Blog post, URL `http://chrisblattman.com/2009/10/15/am-i-actually-sticking-up-for-the-millennium-villages/`.

Blattman, C. (2010). "Evaluating the millennium villages: The saga continues," Blog post, URL `http://chrisblattman.com/2010/10/28/evaluating-the-millennium-villages-the-saga-continues/`.

Blattman, C., N. Fiala, and S. Martinez (2013). "Generating skilled self-employment in developing countries: Experimental evidence from uganda," Quarterly Journal of Economics, Forthcoming.

Buckland, S. T. and P. H. Garthwaite (1991). "Quantifying precision of mark-recapture estimates using the bootstrap and related methods," *Biometrics*, 47, 255–268.

Bump, J. B., M. A. Clemens, G. Demombynes, and L. Haddad (2012). "Concerns about the millennium villages project report," *The Lancet*, 379, 1945.

Butler, D. (2012). "Poverty project opens to scrutiny," *Nature*, 486, 165–166.

Castledine, B. J. (1981). "A bayesian analysis of multiple-recapture sampling for a closed population," *Biometrika*, 67, 197–210.

Catterall, J. S. (1985). "Economic evaluation of public programs," *New directions for program evaluation*, 1985, 99–103.

Chao, A. and P. K. Tsay (1998). "A sample coverage approach to multiple-system estimation with application to census undercount," *Journal of the American Statistical Association*, 93, 283–293.

Chao, A., P. K. Tsay, S. H. Lin, W. Y. Shau, and D. Y. Chao (2001). "Tutorial in biostatistics: The applications of capture-recapture models to epidemiological data," *Statistics in Medicine*, 20, 3123–3157.

Clemens, M. A. and G. Demombynes (2011). "When does rigorous impact evaluation make a difference? the case of the millennium villages," *Journal of Development Effectiveness*, 3, 305–339.

Clemens, M. A. and G. Demombynes (2013). "The new transparency in development economics: Lessons from the millennium villages controversy," Working Paper 342, Center for Global Development, Washington DC.

Clemens, M. A., G. Demombynes, C. Kenny, S. Minard, J. Naudet, and R. Peccoud (2012). "The collision of development goals and impact evaluation," URL `http://www.afd.fr/webdav/shared/PORTAILS/PUBLICATIONS/EUDN/EUDN2012/interventions/Article-Michael-CLEMENS.pdf`, working paper.

Clingingsmith, D., A. I. Khwaja, and M. Kremer (2009). "Estimating the impact of the hajj: Religion and tolerance in islam's global gathering," *The Quarterly Journal of Economics*, 124, 1133–1170, URL `http://scholar.harvard.edu/files/kremer/files/hajj_qje_2009_august.pdf`.

Cohen, J. and P. Dupas (2010). "Free distribution or cost-sharing? evidence from a randomized malaria prevention experiment," *The Quarterly Journal of Economics*, 125, 1–45, URL `http://www.stanford.edu/~pdupas/CohenDupas.pdf`.

Cormack, R. M. (1992). "Interval estimation for mark-recapture studies of closed population," *Biometrics*, 48, 567–576.

Coull, B. A. and A. Agresti (1999). "The use of mixed logit models to reflect heterogeneity in capture-recapture studies," *Biometrics*, 55, 294–301.

Cox, D. R. (1972). "Regression models and life-tables," *Journal of the Royal Statistical Society, Series B*, 34, 187–220.

Darroch, J. N., S. Fienberg, G. Glonek, and B. Junker (1993). "A three-sample multiple-recapture approach to census population estimation with heterogeneous catchability," *Journal of the American Statistical Association*, 88, 1137–1148.

Darroch, J. N., S. L. Lauritzen, and T. P. Speed (1980). "Markov fields and log linear models for contingency tables," *Annals of Statistics*, 8, 522–539.

Datta, G., B. Day, and T. Maiti (1998). "Multivariate bayesian small area estimation: Application to survey and satellite data," *Sankhya, Series A*, 60, 1–19.

Datta, G., M. Ghosh, N. Nangia, and K. Natarajan (????). "Estimation of median income of four-person families: A bayesian approach," in W. Berry, K. Chaloner, and J. Geweke, eds., *Bayesian Analysis in Statistics and Econometrics*, New York, NY: Wiley, 129–140.

Davy, A., K. McPhail, and F. S. Moreno (1999). "BPXC's Operations in Casanare, Colombia: Factoring social concerns into development decisionmaking," Technical Report 31, World Bank.

Dawid, A. P. and S. L. Lauritzen (1993). "Hyper-markov laws in the statistical analysis of decomposable graphical models," *Annals of Statistics*, 21, 1272–1317.

Dehejia, R. H. (2005). "Practical propensity score matching: a reply to Smith and Todd," *Journal of Econometrics*, 125, 355–364.

Dehejia, R. H. and S. Wahba (1999). "Causal effects in nonexperimental studies: reevaluating the evaluation of training programs," *Journal of the American Statistical Association*, 94, 1053–1062.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society, Series B*, 39, 1–22.

DeSouza, C. M. (1992). "An appropriate bivariate bayesian method for analysing small frequencies," *Biometrics*, 48, 1113–1130.

Dominici, F. (2000). "Combining contingency tables with missing dimensions," *Biometrics*, 56, 546–553.

Donner, A. and N. Klar (2004). "Pitfalls and controversies in cluster randomization trials," *American Journal of Public Health*, 94, 416–422.

Duflo, E., R. Glennerster, and M. Kremer (2008). "Using randomization in development economics research: a toolkit," in T. P. Schultz and J. Strauss, eds., *Handbook of Development Economics*, volume 4, North Holland, 3895–3962.

E Stanghellini, P. G. M. v. d. H. (2004). "A multiple-record systems estimation method that takes observed and unobserved heterogeneity into account," *Biometrics*, 60, 510–516.

Elbers, C., J. O. Lanjouw, and P. Lanjouw (2003). "Micro-level estimation of poverty and inequality," *Econometrica*, 71, 355–364.

Fay, R. and R. Herriot (1979). "Estimates of income for small places: an application of james-stein procedures to census data," *Journal of the American Statistical Association*, 366, 269–277.

Feller, A. and A. Gelman (2014). "Hierarchical models for causal effects," URL `http://www.stat.columbia.edu/~gelman/research/published/HierarchicalCausal.pdf`, working paper.

Fienberg, S. E. (1972). "The multiple recapture census for closed populations and incomplete $2k$ contingency tables," *Biometrika*, 59, 591–603.

Fienberg, S. E. (1992). "Bibliography on capture-recapture modelling with application to census undercount adjustment," *Survey Methodology*, 18, 143–154.

Fienberg, S. E. (2000). "Contingency tables and log-linear models: Basic results and new developments," *Journal of the American Statistical Association*, 95, 643–647.

Fienberg, S. E., M. S. Johnson, and B. W. Junker (1999). "Classical multilevel and bayesian approaches to population size estimation using multiple lists," *Journal of the Royal Statistical Society A*, 162, 383–405.

Filmer, D. and L. H. Pritchett (2001). "Estimating wealth effects without expenditure data - or tears: An application to educational enrollments in states of india," *Demography*, 38, 115–132.

for Disease Monitoring, I. W. G. and Forecasting. (1995). "Capture-recapture and multiple-record systems estimation i: History and theoretical development." *American Journal of Epidemiology*, 142, 1047–1058.

Fujii, T. (2005). "Micro-level estimation of child malnutrition indicators and its application in cambodia," Working Paper 3662, World Bank.

Gelman, A. (2006). "Multilevel (hierarchical) modeling: What it can and cannot do," *Technometrics*, 48, 432–435.

Gelman, A. and J. Carlin (2013). "Beyond power calculations to a broader design analysis, prospective or retrospective, using external information," Working paper.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2014). *Bayesian Data Analysis*, Chapman & Hall/CRC texts in statistical science, third edition.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin (2003). *Bayesian Data Analysis*, Chapman & Hall/CRC texts in statistical science, second edition.

Gelman, A. and J. L. Hill (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*, New York, NY: Cambridge University Press.

Gelman, A., J. L. Hill, and M. Yajima (2012). "Why we (usually) don't have to worry about multiple comparisons," *Journal of Research on Educational Effectiveness*, 5, 189–211.

Gelman, A. and D. K. Park (2008). "Splitting a predictor at the upper quarter or third and the lower quarter or third," *The American Statistician*, 62, 1–8.

George, E. I. and R. E. McCulloch (1993). "Variable selection via gibbs sampling," *Journal of the American Statistical Association*, 88, 881–889.

George, E. I. and C. P. Robert (1992). "Capture-recapture estimation via gibbs sampling," *Biometrika*, 79, 677–683.

Ghosh, M. and K. Natarajan (1999). "Small area estimation: A bayesian perspective," in S. Ghosh and M. Dekker, eds., *Multivariate Analysis, Design of Experiments and Survey Sampling*, New York, NY: Wiley, 69–92.

Ghosh, M. and J. N. K. Rao (1994). "Small area estimation: An appraisal," *Statistical Science*, 9, 55–76.

Glonek, G. F. V. (1996). "A class of regression models for multivariate categorical responses," *Biometrika*, 83, 15–28.

Glonek, G. F. V. and P. McCullagh (1995). "Multivariate logistic models," *Journal of the Royal Statistical Society, Series B: Methodological*, 57, 533–546.

Green, D. P., S. E. Ha, and J. G. Bullock (2010). "Enough already about "black box" experiments: Studying mediation is more difficult than most scholars suppose," *The ANNALS of the American Academy of Political and Social Science*, 628, 200–208.

Greenland, S., J. M. Robins, and J. Pearl (1999). "Confounding and collapsibility in causal inference," *Statistical Science*, 14, 29–46.

Grieve, R., J. Cairns, and S. G. Thompson (2009). "Improving costing methods in multicentre economic evaluation: the use of multiple imputation for unit costs," *Health Economics*, 10, 1532.

Group, U. N. D. (2003). "Indicators for monitoring the millennium development goals: Definitions, rationale, concepts and sources." Technical Report ST/ESA/STAT/SER.F/95, United Nations, New York, NY.

Guba, E. G. and Y. S. Lincoln (1989). *Fourth generation evaluation.*, Newbury Park, CA: Sage Publications.

Guberek, T., D. Guzman, M. Price, K. Lum, and P. Ball (2010). "Lethal violations in casanare," A Report by the Benetech Human Rights Program.

Habicht, J. P., C. G. Victora, and J. P. Vaughan (1999). "Evaluation designs for adequacy, plausibility and probability of public health programme performance and impact," *International Epidemiological Association*, 28, 10–18.

Haushofer, J. and J. Shapiro (2013). "Household response to income changes: Evidence from an unconditional cash transfer program in kenya," Working paper.

HemoCue Worldwide (2014). *HemoCue*, Angelholm, Sweden: Radiometer Group, URL `http://www.hemocue.com/en/products/hb-301-kit`.

Hill, J. L. and M. Scott (2009). "Comment: The essential role of pair matching," *Statistical Science*, 24, 54–58.

Ho, D. E., K. Imai, and G. King (2007). "Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference," *Political Analysis*, 15, 199–236.

Humphreys, M., R. S. de la Sierra, and P. van der Windt (2013). "Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration," *Political Analysis*, 21, 1–20.

Hutchinson, B. G. (1969). "The economic evaluation of urban transportation investments," Technical report, London Centre for Environmental Studies.

Ibrahim, J. G., M. H. Chen, and D. Sinha (2001). *Bayesian survival analysis*, Springer series in statistics, 233 Spring St., New York, NY 10013, USA: Springer Science+Business Media, Inc.

Imbens, G. W. (2003). "Sensitivity to exogeneity assumptions in program evaluation," *The American Economic Review*, 93, 126–132.

Imbens, G. W. and D. B. Rubin (2014). *Causal Inference in Statistics and Social Sciences*, Draft.

Imbens, G. W. and J. M. Wooldridge (2009). "Recent developments in the econometrics of program evaluation," *Journal of Economic Literature*, 47, 5–86.

Innovations for Poverty Action (????). "Ultra poor graduation program," URL `http://www.poverty-action.org/ultrapoor`.

ITAD (2013). "Impact evaluation of a new millennium village in northern ghana: Initial design document," Technical report, UK Department for International Development.

Jeffreys, H. (1961). *Theory of Probability*, Clarendon Press, 3 edition.

Jiang, J. and P. Lahiri (2006). "Mixed model prediction and small area estimation," *Test*, 15, 1–96.

Johnson, F. A., H. Chandra, J. J. Brown, and S. S. Padmadas (2010). "District-level estimates of institutional births in ghana: Application of small area estimation technique using census and dhs data," *Journal of Official Statistics*, 26, 341–359.

Kifle, H., M. Hussain, and H. Mekonnen (2002). "Achieving the millennium development goals in africa: Progress, prospects, and policy implications," Technical report, African Development Bank, URL `http://www.cpahq.org/cpahq/cpadocs/Achieving%20the%20MDGs%20in%20Africa.pdf`.

Kreif, N., R. Grieve, R. Radice, and J. S. Sekhon (2011). "Regression-adjusted matching and double-robust methods for estimating average treatment effects in health economic evaluation," URL http://www.lshtm.ac.uk/php/hsrp/reducing-selection-bias/output/regression_adjusted_matching_and_double_robust_methods.pdf, paper presented at the Causal Inference Group Meeting at the Harvard School of Public Health.

Kremer, M. and A. Holla (????). "Improving education in the developing world: What have we learned from randomized evaluations?" *Annual Review of Economics*, 513–542, URL http://scholar.harvard.edu/files/kremer/files/annual_review_kremer_holla_2009.pdf.

Lang, J. B. (2004). "Multinomial-poisson homogeneous models for contingency tables," *Annals of Statistics*, 32, 340–383.

Lang, J. B. (2005). "Homogeneous linear predictor models for contingency tables," *Journal of the American Statistical Association*, 100, 121–134.

Li, F. and A. M. Zaslavsky (2010). "Using a short screening scale for small-area estimation of mental illness prevalence for schools," *Journal of the American Statistical Association*, 105, 1323–1332.

Liu, K. and X. L. Meng (2014). "Comment: A fruitful resolution to simpson's paradox via multiresolution inference," *The American Statistician*, 68, 17–29.

LME4 Authors (2013). "lme4: Linear mixed-effects models using eigen and s4," URL http://cran.r-project.org/web/packages/lme4/index.html.

Lum, K., M. Price, T. Guberek, and P. Ball (2010). "Measuring elusive populations with bayesian model averaging for multiple systems estimation: A case study on lethal violations in casanare, 1998-2007," *Statistics, Politics, and Policy*, 1.

Madigan, D. and J. C. York (1997). "Bayesian methods for estimation of the size of a closed population," *Biometrika*, 84, 19–31.

Madigan, D., J. C. York, and D. Allard (1995). "Bayesian graphical models for discrete data," *International Statistical Review*, 63, 215–232.

Mansour, S., D. Martin, and J. Wright (2012). "Problems of spatial linkage of a geo-referenced demographic and health survey (dhs) dataset to a population census: A case study of egypt," *Computers, Environment and Urban Systems*, 36, 350–358.

McArthur, J. W., P. M. Pronyk, and J. D. Sachs (2011). "Designing, implementing and evaluating complex, goal-oriented adaptive interventions in the millennium villages," URL http://www.csae.ox.ac.uk/conferences/2011-EdiA/plenaries/csae-conf2011-panel2-McArthur.pdf.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models*, London: Chapman & Hall.

McKenzie, D. (2012). "Beyond baseline and follow-up: The case for more t in experiments," *Journal of Development Economics*, 99, 210–221.

Measure DHS/ICF International (2012). "Sampling and household listing manual: Demographic and health surveys methodology," Technical report, Measure DHS, URL http://www.measuredhs.com/pubs/pdf/DHSM4/DHS6_Sampling_Manual_Sept2012_DHSM4.pdf.

Meng, X. L. and A. M. Zaslavsky (2002). "Single observation unbiased priors," *Annals of Statistics*, 30, 1345–1375.

Michelson, H., M. Muniz, and K. DeRosa (2013). "Measuring socio-economic status in the millennium villages: The role of asset index choice," *The Journal of Development Studies*, 49, 917–935.

Mitchell, S., A. Ozonoff, A. M. Zaslavsky, B. Hedt-Gauthier, K. Lum, and B. A. Coull (2013). "A comparison of marginal and conditional models for capture-recapture data with application to human rights violations data," *Biometrics*, 69, 1022–1032.

Molenberghs, G. and E. Lesaffre (1999). "Marginal modelling of multivariate categorical data," *Statistics in Medicine*, 18, 2237–2255.

Muniz, M., B. Nemser, P. M. Pronyk, and E. Quintana (2011). "Survey enumeration manual: Guidelines for enumerators, field supervisors, and data managers," Technical report, Millennium Villages Project, New York, NY, URL https://ciesin.columbia.edu/confluence/download/attachments/91488269/MVP_Y5_Enumeration_Manual.pdf.

MVP (2009). "Study protocol, integrating the delivery of health and development interventions: assessing the impact on child survival in sub-saharan africa." Available from: https://ciesin.columbia.edu/confluence/download/attachments/91488269/MVP+Evaluation+Protocol.pdf.

Nadram, B. (2000). "Bayesian generalized linear models for inference about small areas," in D. Rey, S. K. Ghosh, and B. K. Mallick, eds., *Generalized Linear Models*, Boca Raton: CRC Press, 89–109.

Nature editorial (2012). "With transparency comes trust," *Nature*, 485, URL http://www.nature.com/nature/journal/v485/n7397/full/485147a.html.

Norris, J. L. and K. H. Pollock (1996). "Including model uncertainty in estimating variances in multiple capture studies," *Environmental and Ecological Statistics*, 3, 235–244.

Oakley, A., V. Strange, J. Stephenson, S. Forrest, H. Monteiro, and RIPPLE Study Team (2004). "Evaluating processes: A case study of a randomized controlled trial of sex education," *Evaluation*, 10, 440–462.

O'Brien, P. C. (1984). "Procedures for comparing samples with multiple endpoints," *Biometrics*, 40, 1079–1087.

Pronyk, P. M., M. Muniz, B. Nemser, M. A. Somers, L. McClellan, C. A. Palm, U. K. Huynh, Y. B. Amor, B. Begashaw, J. W. McArthur, A. Niang, S. E. Sachs, P. Singh, A. Teklehaimanot, and J. D. Sachs (2012). "The effect of an integrated multisector model for achieving the millennium development goals and improving child survival in rural sub-saharan africa: a non-randomised controlled assessment," *The Lancet*, 379, 2179–2188.

Pushpangadan, P. (1997). *Conservation and economic evaluation of biodiversity*, India: Science Publishers Inc.

QSR International (2008). *NVivo Qualitative Data Analysis Software: Version 8.*, Cambridge, MA, USA: QSR International.

R Development Core Team (2014). "The r project for statistical computing," URL `http://www.r-project.org/`.

Raghunathan, T. E., D. Xie, N. Schenker, V. L. Parsons, W. W. Davis, E. J. Feuer, and K. W. Dodd (2007). "Combining information from two surveys to estimate county-level prevalence rates of cancer risk factors and screening," *Journal of the American Statistical Association*, 102, 474–486.

Rahman, M. L. and M. F. Alam (1987). *An economic evaluation of some credit programmes designed for the small farmers and landless poor in Bangladesh. Dhaka*, Bangladesh Agricultural University: Bureau of Socioeconomic Research and Training.

Rao, J. N. K. (2003). *Small Area Estimation*, Hoboken, New Jersey: John Wiley and Sons.

Roberts, G. O., A. Gelman, and W. R. Gilks (1994). "Weak convergence and optimal scaling of random walk metropolis algorithms," Technical report, University of Cambridge.

Roberts, H. V. (1967). "Informative stopping rules and inferences about population size," *Journal of the American Statistical Association*, 62, 763–775.

Robins, J. M. and A. Rotnitzky (2001). "Comment on the Bickel and Kwon article, "On double robustness."," *Statistica Sinica*, 11, 920–936.

Robins, J. M., A. Rotnitzky, and M. J. V. der Laan (2000). "Comment on the Murphy and Van der Vaart article, "On profile likelihood."," *Journal of the American Statistical Association*, 95, 431–435.

Rosenbaum, P. R. (1984). "The consequences of adjustment for a concomitant variable that has been affected by the treatment," *Journal of the Royal Statistical Society A*, 147, 656–666.

Rosenbaum, P. R. (2005). "Sensitivity analysis in observational studies," in B. S. Everitt and D. C. Howell, eds., *Encyclopedia of Statistics in Behavioral Science*, volume 4, Chichester, England: John Wiley & Sons, Ltd, 1809–1814.

Rosenbaum, P. R. and D. B. Rubin (1983a). "Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome," *Journal of the Royal Statistical Society, Series B*, 45, 212–218.

Rosenbaum, P. R. and D. B. Rubin (1983b). "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70, 41–55.

Rosyton, P., D. Altman, and W. Sauerbrei (2006). "Dichotomizing continuous predictors in multiple regression: A bad idea," *Statistics in Medicine*, 25, 127–141.

Rubin, D. (2008). "For objective causal inference, design trumps analysis," *The Annals of Applied Statistics*, 2, 808–840.

Rubin, D. B. (1973). "The use of matched sampling and regression adjustment to remove bias in observational studies," *Biometrics*, 29, 185–203.

Rubin, D. B. (1976). "Inference and missing data," *Biometrika*, 63, 581–592.

Rubin, D. B. (1978). "Bayesian inference for causal effects: The role of randomization," *Annals of Statistics*, 6, 34–58.

Rubin, D. B. (1984). "Bayesianly justifiable and relevant frequency calculations for the applied statistician," *The Annals of Statistics*, 12, 1151–1172.

Rubin, D. B. and N. Thomas (2000). "Combining propensity score matching with additional adjustments for prognostic covariates," *Journal of the American Statistical Association*, 95, 573–585.

Rutstein, S. O. and G. Rojas (2006). "Guide to DHS Statistics," DHS toolkit, Demographic and Health Surveys, Demographic and Health Surveys, Calverton, Maryland, URL `http://dhsprogram.com/pubs/pdf/DHSG1/Guide_to_DHS_Statistics_29Oct2012_DHSG1.pdf`.

Sachs, J. (2007). "Rapid victories against extreme poverty," *Scientific American*, 296, 34.

Sachs, J. D. and J. W. McArthur (2005). "The millennium project: a plan for meeting the millennium development goals," *The Lancet*, 365, 347–353.

Sachs, J. D., J. W. McArthur, G. Schidt-Traub, M. Kruk, C. Bahadur, and G. McCord (2004). "Ending africa's poverty trap," *Brookings Papers on Economic Activity*, 1, 117–240, URL `http://www.unmillenniumproject.org/documents/BPEAEndingAfricasPovertyTrapFINAL.pdf`.

Sanchez, P., C. Palm, J. D. Sachs, G. Denning, R. Flor, R. Harawa, B. Jama, T. Kiflemariam, B. Konecky, R. Kozar, E. Lelerai, A. Malik, P. Mutuo, A. Niang, H. Okoth, F. Place, S. E. Sachs, A. Said, D. Siriri, A. Teklehaimanot, K. Wang, J. Wangila, and C. Zamba (2007). "The african millennium villages," *Proceedings of the National Academy of Sciences*, 104, 6775–80.

Sanchez, P., M. S. Swaminathan, P. Dobie, and N. Yuksel (2005). "Halving hunger: it can be done," Technical report, UN Millennium Project Task Force on Hunger.

Schofield, H. (2014). "The economic costs of low caloric intake: Evidence from india," URL `http://scholar.harvard.edu/files/hschofield/files/schofield_calories_and_productivity_2014.01.27.pdf`, working paper.

Schulenburg, J. M. (2000). *The influence of economic evaluation studies on health care decision-making: a European survey*, Amsterdam, Holland: IOS Press.

Shadish, W. R., M. H. Clark, and P. M. Steiner (2008). "Can nonrandomised experiments yield accurate answers? a randomized experiment comparing random and nonrandom assignments," *Journal of the American Statistical Association*, 103, 1334–1343.

Shiell, A., P. Hawe, and L. Gold (2008). "Complex interventions or complex systems? implications for health economic evaluation," *British Medical Journal*, 336, 1281–3.

Simler, K. R. (2006). "Nutrition mapping in tanzania: An exploratory analysis," FCND Discussion Paper 204, International Food Policy Research Institute, 2033 K Street, NW, Washington DC.

Smith, P. J. (1991). "Bayesian analyses for a multiple capture-recapture model," *Biometrika*, 78, 399–407.

Sorenson, G., K. Emmons, M. K. Hunt, and D. Johnston (1998). "Implications of the results of community intervention trials," *Annual Review of Public Health*, 19, 379–416.

Stan Development Team (2013). "Stan: A c++ library for probability and sampling, version 1.3," URL `http://mc-stan.org/`.

Starobin, P. (2013). "Does it take a village?" URL `http://www.foreignpolicy.com/articles/2013/06/24/does_it_take_a_village`.

StataCorp (2011). *Stata Statistical Software: Release 12*, College Station, Texas: StataCorp LP.

Stuart, E. A. and A. M. Zaslavsky (2005). "Using administrative records to predict residency: Final report," *U.S. Bureau of the Census*.

Sutherland, J. M., C. J. Schwarz, and L. P. Rivest (2007). "Multilist population estimation with incomplete and partial stratification," *Biometrics*, 63, 910–916.

The Economist (2012). "Millennium bugs: Jeffrey sachs and the millennium villages," URL http://www.economist.com/blogs/feastandfamine/2012/05/jeffrey-sachs-and-millennium-villages.

Tsay, P. K. and A. Chao (2001). "Population size estimation for capture-recapture models with applications to epidemiological data," *Journal of Applied Statistics*, 28, 25–36.

UN Millennium Project (2005). "Investing in development: A practical plan to achieve the milliennium development goals," Technical report, UN Millennium Project, New York, URL http://www.unmillenniumproject.org/reports/fullreport.htm.

UNESCO Institute for Statistics (2010). "Global education digest 2010: Comparing education statistics across the world," Technical report, UNESCO Institute for Statistics, Montreal, Quebec, URL http://www.uis.unesco.org/Library/Documents/GED_2010_EN.pdf.

United Nations, G. A. (2000). "55/2. united nations millennium declaration," available from http://www.un.org/millennium/declaration/ares552e.htm.

US Census Bureau (2013). *CSPro: Census and Survey Processing System, Version 5.0.3*, Washington DC, USA: US Census Bureau, Macro International, and Serpo, S.A., URL http://www.census.gov/population/international/software/cspro/.

van der Heijden, P. G. M., E. Zwane, and D. Hessen (2009). "Structurally missing data problems in multiple list capture–recapture data," *Advances in Statistical Analysis*, 93, 5–21.

Wanjala, B. M. and R. Muradian (2013). "Can big push interventions take small-scale farmers out of poverty? insights from the sauri millennium village in kenya," *World Development*, 45, 147–160.

Wordsworth, S., A. Ludbrook, F. Caskey, and A. Macleod (2005). "Collecting unit cost data in multicentre studies: Creating comparable methods," *The European Journal of Health Economics*, 6, 38–44.

World Health Organization (2003). "A summary of the findings of the commission on macroeconomics and health," Technical report, World Health Organization CMH support unit.

You, Y. and Q. M. Zhou (2011). "Hierarchical bayes small area estimation under a spatial model with application to health survey data," *Survey Methodology*, 37, 25–37.

Zaslavsky, A. M. (2011). "Sampling from a bayesian menu," *Statistical Science*, 26, 235–237.

Zaslavsky, A. M. and G. S. Wolfgang (1990). "Triple-system modeling of census, post-enumeration survey, and administrative list data," *Proceedings of the Survey Research Section, American Statistical Association*, 668–673.

Zaslavsky, A. M. and G. S. Wolfgang (1993). "Triple-system modeling of census, post-enumeration survey, and administrative list data," *Journal of Business and Economic Statistics*, 11, 279–288.

Zwane, E. N., K. van der Pal-de Bruin, and P. G. M. van der Heijden (2004). "The multiple-record systems estimator when registrations refer to different but overlapping populations," *Statistics in Medicine*, 23, 2267–2281.